Functional and Molecular Diversity of the Diatom Family Leptocylindraceae

Thesis

THE OPEN UNIVERSITY, LONDON
STAZIONE ZOOLOGICA ANTON DOHRN, NAPLES

Functional and Molecular Diversity of the Diatom Family Leptocylindraceae

Thesis submitted for the degree of Doctor in Philosophy (Ph.D.) in Life and Biomolecular Sciences

Aikaterini Pargana

9/30/2016

Director of Studies: Dr Adriana Zingone

Laboratory of Ecology and Evolution of Plankton

Stazione Zoologica Anton Dohrn, Naples, Italy

Internal Supervisor: Dr Maria Immacolata Ferrante

Laboratory of Ecology and Evolution of Plankton

Stazione Zoologica Anton Dohrn, Naples, Italy

External Supervisor: Dr Chris Bowler

Plant and Diatom Genomics Laboratory,

Institut de Biologie de l'École Normale Supérieure, Paris, France

Contents

Abstract

The focus of this PhD project is the functional and molecular diversity of Leptocylindraceae diatom species, the study of which can lead to a better understanding of long standing questions regarding the ecology and evolution of phytoplankton. A wide range of tools, spanning from microscopical observations and physiological measurements to molecular techniques and high throughput sequencing, is utilized during this attempt. The genus *Leptocylindrus* has been chosen as the main target species due its worldwide and at the same time local importance in the Gulf of Naples and also because of the already extended study of the species in Stazione Zoologica Anton Dohrn (SZN) towards the direction mentioned above.

Leptocylindraceae are centric diatoms that occupy a basal position in the diatom phylogeny and are abundant in marine plankton worldwide. In the Gulf of Naples (GoN), five out of the six species are found; *L. minimus* is known to be absent from the Mediterranean environment. The family shows a morphological conservation but the seasonal patterns between the species differ considerably. Indeed, physiological experiments of two *Leptocylindrus* species that show contrasting seasonality confirmed their opposed preferences regarding temperature as well as a high intraspecific phenological variability. In addition, the analysis of transcriptomes acquired for the three temperatures of one of them (*L. aporus*) indicated the possibly important role of transposable elements in response to stress and diatom adaptation. Furthermore, the transcriptomes of all *Leptocylindrus* species were explored in order to detect basic intra- and interspecific similarities and/or differences. HTS sequencing data from the MareChiara station in GoN and from the Tara expedition in the world's seas were analyzed in order to assess the actual diversity of this important diatom family. A significant level of intraspecific variability was detected while the distribution of species and populations at spatial and temporal scale supported the functional differences among them that account for their distinct seasonality and their adaptation to different environmental conditions.

Acknowledgements

This study was carried out at the Department of Integrated Marine Ecology (IME) at Stazione Zoologica “Anton Dohrn” in Napoli, Italy. I would like to thank my supervisor, Adriana Zingone, for the guidance and support during the work and writing of my thesis. Her comments and revision of the language and structure as well as her advices on scientific thinking were essential for the final form of this thesis. I want also to stress the important contribution of my internal supervisor, Mariella Ferrante, and my external supervisor, Chris Bowler, to the completion of the project. Their expertise in the related fields and their advices were a valuable assistance.

I am grateful to all the members of the lab for all the inspiring discussions we had but I would specifically like to thank Roberta Piredda and Maria Paola Tomasino for providing me data, tools, related guidance and useful comments but also for being my friends and making the working days pleasant. I want to thank Chetan Gaonkar for helping me in the lab and always supporting me during tough times but also for introducing me to the amazing Indian tradition and cuisine. I could not ask for a better person to share my office and the three years of my Ph.D. with. I would like to thank Carmen Minucci, Alessandro Manfredonia, Ferdinando Tramontano, Elvira Mauriello, Marco Borra, Raimondo Pannone, Elio Biffali, Flora Palumbo for technical assistance in parts of this work as well as for their help in laboratory work. I want to thank all the people I met in Stazione and proved once more that friendship can make things in life (and work!) easier and brighter: Laura V., Laura E., Eleonora, Solenn, Arianna, Greta, Angela, Davide, Mariano S., Vincenzo, Mariano A., Romain, Luigi, Ida, Nico and Gauri.

I wish to dedicate this work to my family and my friends back home, in Greece. Without their love and psychological support I would never be able to finish this work.

Finally, I would like to thank Italy, especially the particular city of Napoli, and Stazione Zoologica in general for this life changing experience that I will never forget.

List Of Abbreviations

© / ®, TM	Copyright / Registered, unregistered TradeMark
°C	Degree Celcius
ALDH	Aldehyde Dehydrogenase
ANOVA	Analysis Of Variance
ATP	Adenosine Triphosphate
BioMarKs	Biodiversity of Marine Eukaryotes
BLAST	Basic Local Alignment Search Tool
bp	Base pair
ca.	c*irca* (approximately, about)
CCA	Canonical-correlation analysis
CDD	Conserved Domains Database
CPM	Counts Per Million
DCM	Deep Chlorophyll Maximum
DE	Differential expression/ Differentially expressed
DNA	Deoxyribonucleic acid
dNTP	Deoxynucleotide triphosphate
e.g	*Exempli gratia* (for example)
et al.	*et alia* (and other)
etc	*et cetera* (and so on)
FC	Fold Change
FDR	False Discovery Rate
GO	Gene Ontology
GOEA	Gene Ontology Enrichment Analysis
GoN	Gulf of Naples
HCA	Hierarchical Cluster Analysis
HGT	Horizontal Gene Transfer
HSF	Heat Shoch Transcription Factor
HSP	Heat Shock Protein
HTS	High Throughput Sequencing
i.e.	*id est* (that is to say)
L:D	Light:Dark hour photoperiod
LM	Light Microscope
LSU	Large Subunit Ribosomal DNA
LTER-MC	Long Term Ecological Research station – MareChiara
MECA	Area of Management and Ecology of Coastal Areas
ML	Maximum Likelihood
μ_{max}	Maximum Growth Rate
NCBI	National Center for Biotechnology Information
NJ	Neighbour Joining
PCA	Principal Component analysis
OTU	Operational Taxonomic Unit
qRT-PCR	Quantitative Real Time Polymerase Chain Reaction
RNA	Ribonucleic acid
RNA-seq	RNA sequencing
rRNA	Ribosomal RNA
SNP	Single Nucleotide Polymorphism
SYM1	Stress-inducible yeast Mpv17
SZN	Stazione Zoologica Anton Dohrn Napoli
TE	Transposable Element
TUB	Tubulin
UCSC	University of California Santa Cruz

List of Tables

List of Figures

Chapter 1. General Introduction

1.1. Introduction

Oceans are the largest but the least known habitat on earth. Despite the numerous expeditions that have already provided a massive amount of new information about them and their organisms only a first glimpse into the marine ecosystem has been achieved so far. There is a lack of deep understanding of the parts that shape the whole and this is probably because of the size of the marine ecosystem and the difficulties in accessing it. However, about two decades ago, Biological Oceanography and Marine Ecology entered a new era by applying new technology to old questions with one of the most significant examples being the description of the hydrothermal vents discovered on the deep ocean floor (Van Dover, 1990; Tunnicliffe and Fowler, 1996) using genome-based technologies such as proteomics and metagenomics (Markert et al., 2007; Grzymski et al., 2008). Genome-enabled technology has equipped Marine Ecology with an extremely powerful tool which offers us the potential to fill in gaps in our understanding of marine organisms, their evolution and adaptation to the environment as well as the diversity and functioning of marine communities (Mock and Kirkham, 2011). Ecology and genomics might be considered as a mismatched couple due to their contradictory approaches from two completely different viewpoints; ecology is all about observing the interactions of organisms and their communities with the environment and building hypotheses based on them (Odum, 1977; Putman and Wratten, 1984) while in contrast genomics is not primarily driven by hypotheses but by technology and has so far focused on data collection. Recent developments such as high throughput sequencing technologies (e.g. Illumina, 454 GS-Titanium, SOLID) have significantly advanced genomic disciplines (Schuster, 2008) and contributed in a great extent to the generation of new hypotheses. But even though the deductive system of ecology to explain evolution seems not to match the inductive approach per se of genomics, all genome projects with marine organisms so far have led to pioneer insights into their biology and evolution (e.g. Rocap et al., 2003; Armbrust et al., 2004; Bowler et al., 2008; Worden et al. 2009). There is still a lot to be investigated and understood before someone can say that genomic information can be integrated into the general ecological concepts and the main reasons are that a) the identification of single

genes, gene families and their importance for the biology and evolution of organisms is still in the discovery phase and b) the way to select and integrate the sequence data into ecosystem models is yet in a very premature stage (Mock and Kirkham, 2011).

1.2. The marine environment

Oceans are vast and therefore they cannot be homogenous throughout their whole extent. However, depending on the purposes of each independent research or investigation conducted, oceans can be divided into certain categories. For the study of the evolutionary ecology of phytoplankton, Kilham and Kilham (1980) followed the general practice of biological oceanography and categorized all ocean waters based on their nutrient concentrations and temporal variability into a) estuarine, b) coastal and c) oceanic. Estuarine waters are rich in nutrients but highly variable since the interaction of river deposition and tidal currents create an unstable, dynamic environment. Coastal waters are less nutrient rich than estuarine waters because of the nutrient dispersion over wider areas, utilization and precipitation. Temporal variability is a typical feature of coastal environments and is mainly due to seasonal variations in solar radiation and temperature. Temperature is an important factor in marine environments and one of the reasons is that water density varies with temperature. Stratification or mixing of water can be induced by temperature changes and that will have an effect on hydrography and nutrient availability. Despite that, the environmental perturbations caused by seasonal processes in the coastal area are still less extreme and of longer period than in the estuarine environment (Kilham and Kilham, 1980).

Oceanic waters are characterized by low nutrient concentrations and high temporal stability. In this environment, nutrient supply is mainly determined by atmospheric inputs and diffusive transport from deep water (Menzel and Spaeth, 1962). The maintenance of the very low nutrients level results from the very slow transfer, due to density discontinuities, from the nutrient rich deep water to the surface water combined with the incorporation of nutrients into living organisms (Walsh, 1976). The temporal stability does not mean that supply rates are

unchangeable but it implies that changes in nutrient supply are so slow that any increase is directly incorporated into biomass (Kilham and Kilham, 1980).

1.3. Phytoplankton: Diatoms

Phytoplankton is a polyphyletic assemblage of photosynthetic microscopic organisms that live in aquatic environments. They contribute about 40-50% of global primary productivity even though their photosynthetic biomass represents approximately only 0.2% of the one in land (Falkowski et al., 1998). That means that they have a very high biomass turnover compared to land plants and therefore represent a considerably dynamic pool of organic carbon that can be more easily affected by environmental alterations and vice versa (Nelson et al., 1995; Field et al., 1998; Smetacek, 1999; Sayre, 2010). Phytoplankton dynamic growth along with its important contribution to the global carbon cycle makes it a key player in our perception of the global biogeochemical cycles of elements and that is why phytoplanktonic organisms were an early target for marine genome sequencing projects (Mock and Kirkham, 2011). At least one genome sequence from each major algal taxon is available (Table 1.3.1) and due to that we now have a sense of the major forces shaping the evolution of algae.

Table 1.3.1 Algal groups and selected representative organisms whose genome has been sequenced (Parker et al., 2008). The list has been updated to 2013.

Organism	Group	Genome size	Year of completion
Guillardia theta	Cryptomonad	0.551 Mb	2001
Thalassiosira pseudonana	Diatom	34.5 Mb	2004
Cyanidioschyzon merolae	Red alga	16.5 Mb	2004
Ostreococcus tauri	Green alga	12.6 Mb	2006
Aureococcus anophagefferens	Stramenopile	56.7 Mb	2007
Phaeodactylum tricornutum	Diatom	27.4 Mb	2007
Emiliania huxleyi	Haptophyte	165 Mb	2008
Micromonas pusilla	Prasinophyte	21.5 Mb	2008
Volvox carteri	Green alga	131.2 Mb	2010
Bathycoccus prasinos	Green alga	15 Mb	2012
Chondrus crispus	Red alga	105 Mb	2013

Diatoms are among the most common types of phytoplankton and one of the most successful clades of eukaryotic, single-celled photosynthetic organisms in the modern ocean (Smetacek, 1999). Many species, like *Leptocylindrus* species, form chains composed of sister cells with no

cytoplasmatic contact with one another. The trademark of diatoms is an ornate, siliceous cell wall called the frustule which is composed of two halves called thecas; one half fits in the other slightly larger half (Kooistra et al., 2007). Each half consists of two parts, the valve which is the flat surface of the theca and the girdle which is the side wall. The valve belonging to the largest half is called epivalve while the other one that fits in it hypovalve (Round et al., 1990). The valve often has pores (areolae), spines, hyaline areas, processes and other protrusions that might be advantageous for the cell protection against grazers or involved in the increase of buoyancy. The transparency of the frustule allows light to pass through for photosynthesis while its ultrastructure with pores and poroids allows dissolved gases and nutrients to enter and exit. Traditionally, diatoms are divided in two main classes based on the shape and symmetry of their cell walls: centric and pennate diatoms (Fig. 1.3.1). Centric diatoms (*Chaetoceros, Thalassiosira, Leptocylindrus* and *Skeletonema*) are either radially symmetric or bi/multipolar while pennates are asymmetrical (*Pseudo-nitzschia, Phaeodactylum*). Pennates can be further divided based on the presence or not of a raphe.

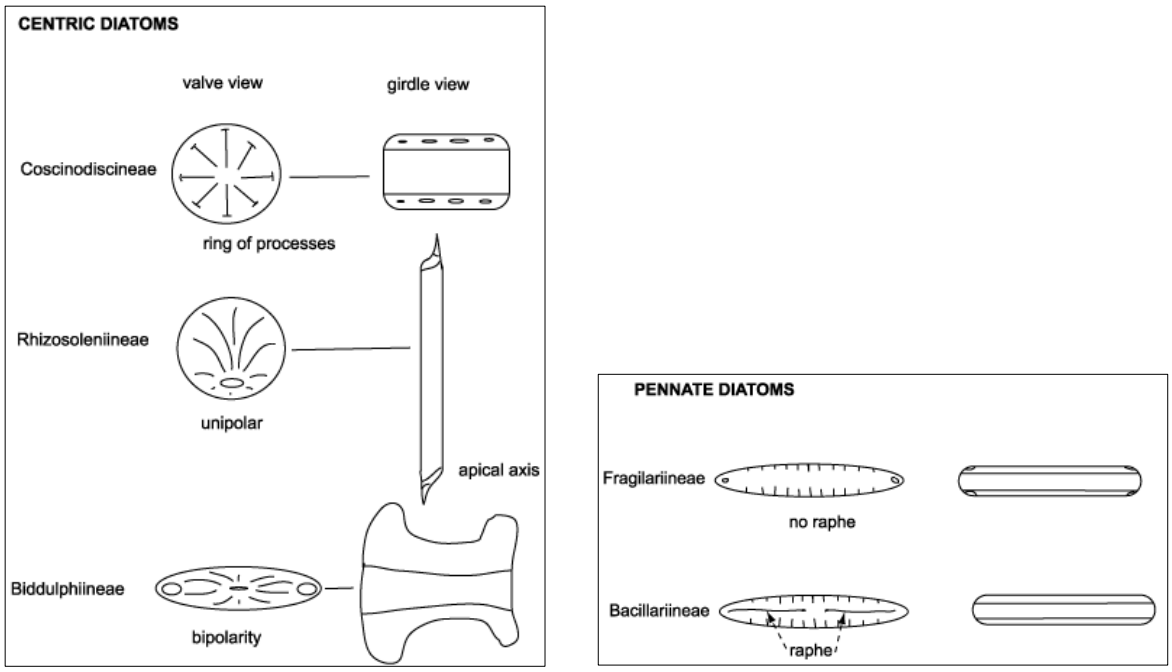


Figure 1.3.1 Schematic diagram of centric and pennate diatom suborders redrawn from Hasle and Syvertsen (1997).

Radial centrics are the most ancient group, whereas bi/multipolar centrics emerged from them and therefore are a newer group. Finally, pennates evolved from the polar centrics (Fig. 1.3.2).

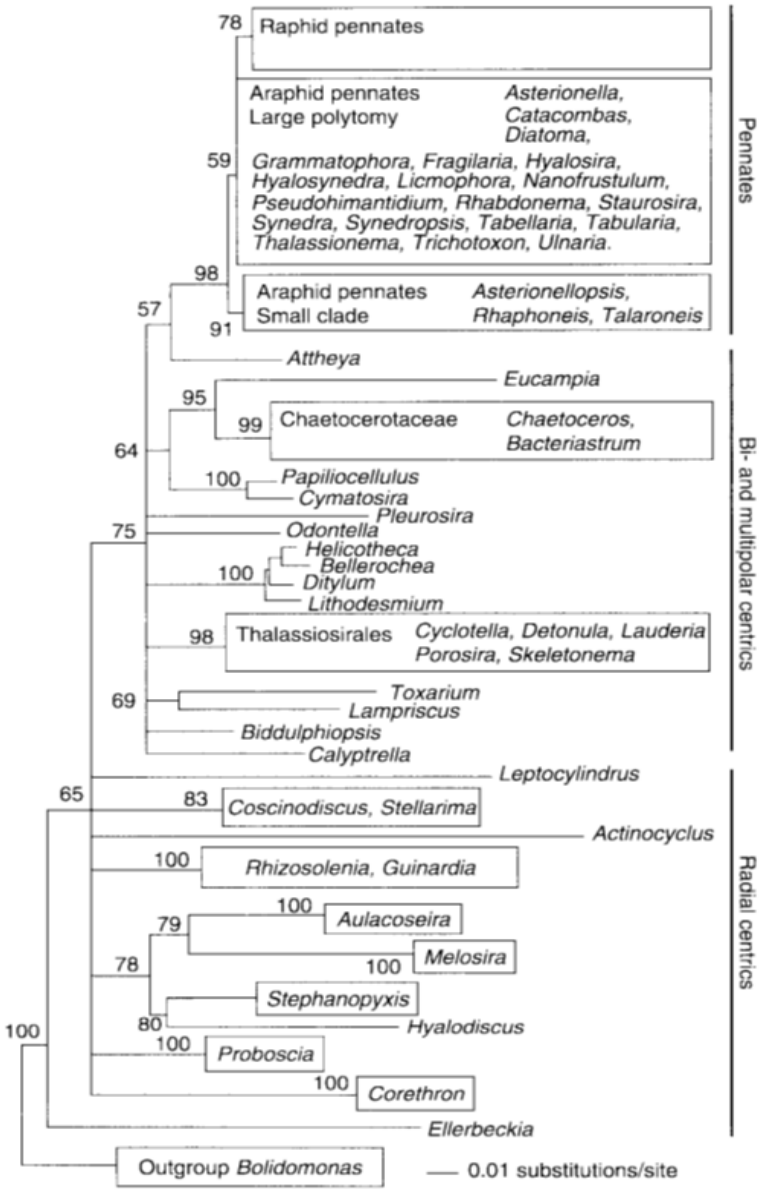


Figure 1.3.2 Neighbour joining (NJ) phylogeny inferred from maximum likelihood pair-wise distance among nuclear SSU rDNA sequences of various diatom genera (Kooistra et al., 2007).

The first critical step in the evolution of algae was the engulfment of a photosynthetic cyanobacterium by a unicellular eukaryote about 1.2 billion years ago. That endosymbiont became the chloroplast and this primary endosymbiosis gave rise to three major photosynthetic lineages: the green, the red and the glaucophyte lineage (Falkowski et al., 2004). All other lineages, including diatoms, were derived from additional endosymbiotic events (secondary and tertiary endosymbiosis) between heterotrophic and autotrophic eukaryotic organisms (Parker et al., 2008). Genome sequencing of *Thalassiosira pseudonana* and *Phaeodactylum tricornutum* provided data that revealed more than 70% of genes derived either from red or green sources are in fact of green lineage origin. According to Bowler et al. (2008), *P. tricornutum* and *T. pseudonana*

have been found to share only 57% of their genes without any synteny (conservation of gene order). In addition to vertical evolution described above, indications for horizontal gene transfer (HGT) between bacteria and diatoms have been found by microscopical and physiological studies (Schmid, 2003). The molecular data that came later confirmed this HGT evolution and gave rise to additional novel features. For example, most of the genes of prokaryotic origin were found to belong to the shell group (moderately common genes in prokaryotes) and their functions are related to mitochondrial metabolism, including one of the most interesting findings, genes encoding proteins from the urea cycle (Bowler et al., 2008). Organisms from green lineage were also detected as gene donors for HGT to diatoms (Moustafa et al., 2009). Armbrust et al. (2004) observed a high degree of polymorphisms in diatom genomes which is probably the result of their complex evolution. Maintaining polymorphic alleles may be favoured by positive selection because it may improve survival in fluctuating ocean environment by providing diatoms with the necessary set of tools when the environment is changing (Mock and Kirkham, 2011). Another possible way to adapt quickly to extreme environments is the re-arrangement of the genome through transposable elements which have been found to contribute 6.4% of the *P. tricornutum* genome (Maumus et al. 2009).

Diatoms are widespread in the plankton and benthos of freshwater, coastal and oceanic habitats and even in temporarily wet terrestrial environments. They often form widespread blooms in temperate and polar seas. In particular, centrics are planktonic with many species constituting major part of phytoplankton blooms in the coastal ecosystems and upwelling regions. On the other hand, pennates are often benthic, found in sediments or as epiphytes on invertebrates or macrophytes; yet, they still can be successful in both the open ocean and turbulent coastal waters (Kooistra et al., 2007). Diatom productivity maintains most of the world's fisheries while their fossilized remains are the main source of petroleum. In fact diatoms account for a strong share of the organic carbon and oxygen produced on Earth (Falkowski et al., 1998; Field et al., 1998). Some of the main characteristics that make diatoms such successful settlers in almost every habitat are their central vacuole that can store nutrients when in favorable conditions for later use, their

light-harvesting system that is able to protect them against high light intensity, their highly efficient CO_2 uptake mechanisms and the formation of resting stages to overcome unfavorable to growth periods (Kooistra et al., 2007).

Diatoms reproduce mostly by cell division, a type of asexual reproduction called vegetative phase during which the overlapping valves of the frustule separate and each secretes a new, smaller half. The most discrete property of diatoms life cycle is this gradual reduction in cell size which can be regained either by vegetative cell enlargement or by sexual reproduction (D'Alelio et al., 2010, Fig. 1.3.3). In the latter case, resistant stages known as auxospores are formed which eventually give rise to larger cells (Drebes, 1977; Chepurnov et al., 2004).

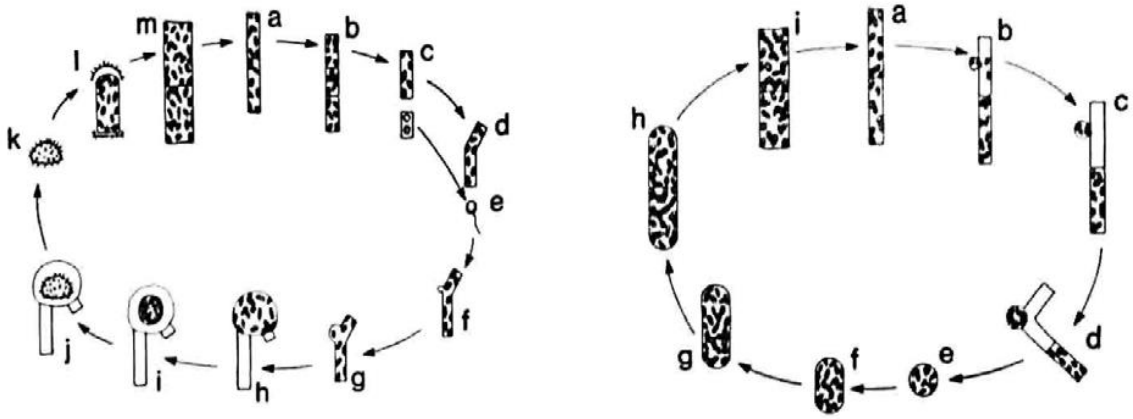


Figure 1.3.3 Sexual reproduction (left) and vegetative cell enlargement (right) in the life cycle of *L. danicus.* In the left cycle, sexuality (b-f) produces auxospore (h) within which the resting spore forms (j-k), later germinating (l-m) to produce cells of maximum diameter. In the right cycle, extrusion of vegetative cell contents usually occurs midway along pervalvar axis (no sexuality, b). When extrusion is completed an auxospore-like structure is formed (c-d) which is detached from the parent cell (e) and finally elongates and germinates cells of maximum diameter (f-g) (French and Hargraves, 1985).

Most species rely on sexual reproduction for their size restoration and therefore diatoms are considered to have a diplontic life history (diploid vegetative cells with haploid gametes) which is unique among algae (Al-Kubaisi, 1981). Sexual reproduction takes place only under two conditions, the first one being the already mentioned reduction under a specific size threshold and the second one being the right environmental and physiological conditions (Geitler, 1932). Under favorable conditions such as high nutrient availability and appropriate temperature and light conditions, rapid reproduction of diatoms occurs (blooms) (Diersing, 2009). Silica, nitrogen and iron are the most critical drivers of the diatom growth in the ocean (Smetacek, 1999; Quigg et

al., 2003). The blooms usually decline because nutrients are depleted or physical factors get less favorable and in that case many diatom species can turn dormant (spore formation) and sink to lower water levels until a new bloom occurs. Diatom cells that do not produce gametes keep dividing mitotically resulting in smaller and smaller progeny until their death. When diatoms die, their glassy frustules settle to the bottom of the sea floor forming thick deposits of siliceous material, known as diatomaceous ooze, huge fossil deposits of which are found inland, mined and used in products such as temperature and sound insulators, for clarifying beer, filters for swimming pools, as mild abrasives in toothpaste etc.

Originally, it was considered that there are approximately 12,000 diatom species but since molecular studies in diatoms became more frequent and sequence data were employed to investigate species-level variation, the estimated number raised to at least 30,000 and probably ca 100,000 species (Mann and Vanormelingen, 2013). Regarding their genetic diversity, diatoms are not distributed homogeneously across the phylogenetic tree. There are genera such as the *Pseudo-nitzschia* Peragallo (Lundholm et al., 2012), *Chaetoceros* Ehrenberg (Rines and Hargraves, 1990; Kooistra et al., 2010) and *Skeletonema* Greville (Sarno et al., 2005, 2007; Zingone et al., 2005; Alverson et al., 2008; Kooistra et al., 2008) that are highly diverse while others like *Leptocylindrus* genus are relatively poor. In the recent years many studies integrating molecular phylogenetics, morphological, ultrastructural and biological information uncovered numerous cases of genetically distinct and even reproductively isolated groups of strains that were otherwise undistinguishable with microscopy resulting in many cases in their designation as cryptic or pseudo-cryptic species (Sarno et al., 2005; Amato et al., 2007; Nanjappa et al., 2013). The species of interest of this study is one of these cases. The high genetic diversity of diatoms is a consequence of different modes of evolution, and hence adaptation to different ecological niches in the marine environment. Each genetically distinct population adapted to specific environmental conditions, called an ecotype, is characterized by a relative fitness which ultimately regulates the distribution and abundance of the species in the ocean. At the same time, most diatom species encounter dramatic temporal fluctuations in environmental conditions and phenotypically buffer

their physiological functioning in order to face this environmental heterogeneity (Bradshaw, 1965; Thompson, 1991; Everroad and Wood, 2012). This so called adaptive phenotypic plasticity is a well-established mechanism in phytoplankton to respond to short-term environmental variability. It is theoretically predicted that fluctuating versus constant environmental conditions will select either for differentially adapted populations (ecotypes) or enhanced plasticity of all – purpose genotypes (Cooper and Lenski, 2010; Alpert and Simms, 2002). The presence of adapted and/or plastic populations was also explored in the current thesis on the case of Leptocylindraceae diatom species and to that end more than one strains of each species were used.

1.4. *Leptocylindrus* species

The diatoms of the *Leptocylindrus* genus are centric diatoms, common in the marine plankton worldwide. According to a recent DNA metabarcoding study based on High Throughput Sequencing (HTS) at six coastal sites in European coastal waters, individual *Leptocylindrus* species are found to be widespread but not all of them are everywhere and their diversity appears to be low (Nanjappa et al., 2014). The genus was found to consist of at least five species instead of two as originally considered (Nanjappa et al., 2013, Table 1.4.1). One of the species was so different from the rest that a new genus, *Tenuicylindrus*, had to be established in order to circumscribe it properly.

Table 1.4.1. Main distinctive morphological characters in *Leptocylindrus* and *Tenuicylindrus* species (French and Hargraves, 1986; Nanjappa et al., 2013).

	L. aporus	*L. convexus*	*L. danicus*	*L. hargravesii*	*L. minimus*	*L. minimus*	*T. belgicus*
Cell diameter (µm)	3.5–10.6	3.0–8.0	3.0–13.0	3.0–15.0	1.5–4.5	2.0–5.2	2.0–2.5
Cell length (µm)	12.5–33.0	22.0–65.0	22.0–75.0	30.0–90.0	–	–	23.6–50.0
Plastid no.	3–13	3–11	7–36	9–55	1–2	1–2	2
Plastid shape	Discoid, ovoid	Ovoid, elongated	Discoid	Discoid	Elongated	Elongated	Elongated
Cells per chain	2–24	2–68	2–165	2–162	–	–	2–14
Valve to mantle ratio	2.9–8.4	2.4–4.0	5.6–11.5	5.3–14.3	–	–	–
Striae (in 1 µm)	9–13	6–10	8–13	7–11	13–18	8–13	7–11
Valve areolae (in 1 µm)	10–14	10–14	18–30	10–14	–	–	12–22
Constriction at the cell junction	Small	Marked	Small	Small	Small	–	Absent
Auxospores and resting spores	Not observed	Not observed	Spiny, semiglobular	Spiny, semiglobular	Spiny, globular with a neck	Not described	Not observed
Sub-central pore	Absent	Absent	Adjacent to annulus	Slightly away from annulus	Absent	Absent	Absent

The reappraisal has been based on frustule and spore morphology and on three nuclear- and three plastid-encoded markers. Eighty-six strains were obtained to assess the morphological and genetic diversity of *Leptocylindrus*. The main distinctive morphological characters in *Leptocylindrus* and *Tenuicylindrus* are shown in Table 1.4.1. Almost all markers contained sufficient information to distinguish the species with 5.8S rDNA and plastid SSU rDNA providing the lowest number of variable positions (33 and 34, respectively) whereas the nuclear SSU rDNA provided the highest (358). All the phylogenies inferred from their alignments showed the same relationships as the one depicted in Fig. 1.4.1.

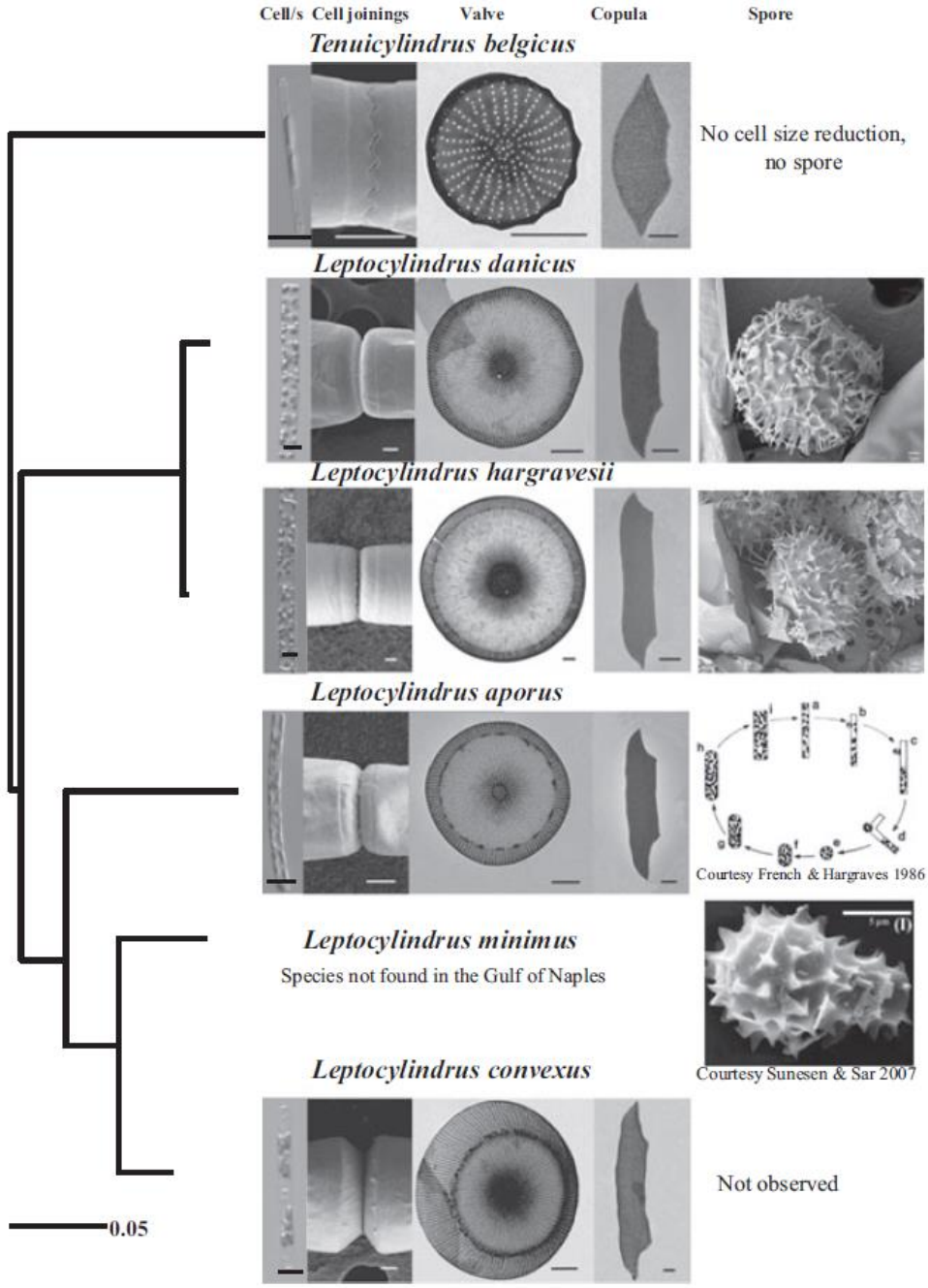


Figure 1.4.1 Differential characters for *Leptocylindrus* and *Tenuicylindrus* species plotted on the SSU rDNA maximum likelihood tree. Scale bar, LM: 10 lm, EM: 1 lm (Nanjappa et al., 2013).

An interesting fact in the case of Leptocylindraceae is that, despite simple morphology retained by all species (narrow, cylindrical cells), they are more distantly related compared to species in the genera *Skeletonema* and *Pseudo-nitzschia* that show higher morphological differentiation among species (Nanjappa et al., 2013). In addition, the seasonal distribution and the life cycle are also quite diverse among species. *L. aporus* is apparently the most abundant species in the Gulf of Naples (GoN), where it was present from mid-July to mid-November; *L. danicus*, which is also relatively abundant, was present from mid-November to mid-July, whereas *L. hargravesii* was

rarely found and only in December and January. *L. convexus* was found from the end of November to the end of July, whereas *L. minimus* was not found at all in the GoN, instead it was *Tenuicylindrus belgicus* that had been misidentified under that name. *T. belgicus* was present from the end of August to the beginning of November (Fig.1.4.2).

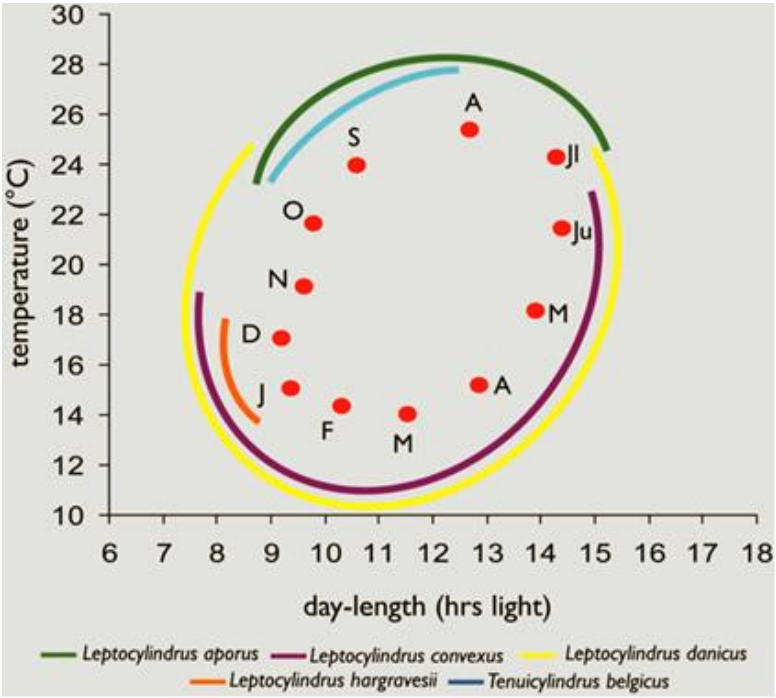


Figure 1.4.2 Seasonal distribution of Leptocylindraceae species at the LTER-MC station (Gulf of Naples, GoN, Mediterranean Sea) based on year-round strain isolations followed by microscopic observations and molecular identifications. Temperature and light conditions related to the season where each species was detected are also depicted for the GoN (Nanjappa et al., 2014b).

Regarding their life cycle, *L. aporus* shows vegetative cell enlargement through "auxospore-like structures", with no ability of spore formation, *L. danicus,* besides vegetative enlargement, shows sexual reproduction and spore formation (Fig. 1.4.1). The resting spores of *L. danicus* derive directly from auxospores, a characteristic only rarely found in diatom species. *L. hargravesii* is quite similar to *L. danicus*, morphologically speaking but also regarding its life cycle. Morphological distinction between these two species is possible only under electron microscopy. In *L. convexus* no resting spore formation has been observed neither cell enlargement but cell size does vary over time. Finally, *T. belgicus* maintains a rather constant cell size, with no indication so far of vegetative enlargement or of auxospore formation. Also spores have not been observed in the latter species. All the information on the seasonal distribution and life cycle of

Leptocylindraceae are mainly based on Nanjappa's (2013) observations on strains cultivated from isolations that were performed almost weekly for more than one year. Despite this exhaustive try for seasonal reconstruction, there is always the possibility that a rare species, such as *L. hargravesii*, escaped isolation during a specific season. In any case, these results made up the first evidence of a higher diversity of the genus *Leptocylindrus*, which was further supported by a functional classification based on oxylipin diversity (Nanjappa et al., 2014b), the results of growth and metabolomics experiments (Nanjappa, 2012) as well as by the first indications deriving from a comparative transcriptomics investigations on *L. danicus* and *L. aporus* (Nanjappa et al., submitted).

In more details, the production and diversity of C_{20} and C_{22} non-volatile oxylipins were analysed in Leptocylindraceae by Nanjappa et al. (2014b). Oxylipins are secondary metabolites implicated in several biological processes such as signaling and defense against biotic stressors and grazers. Lipoxygenases (LOXs) are enzymes involved in oxylipin biosynthesis by oxidizing fatty acids to hydroperoxides. In the Leptocylindraceae, species-specific lipoxygenase activity and oxylipin patterns reflected the phylogenetic relationships observed with molecular markers (Nanjappa et al., 2014b).

To explore if the diversity found in taxonomy and oxylipin patterns was paired with similar degree of diversification at the biochemical and physiological level, metabolite levels and growth characteristics at different temperatures were analysed. Temperature plays an important role in the temporal and spatial distribution of a species since growth is only possible within some relatively limited range (Pörtner and Farrell, 2008; Montes-Hugo et al., 2009). Different temperatures have different effects on the rate of enzyme catalyzed reactions in various metabolic processes of the cells, affecting ultimately their growth and physiology. Therefore, Nanjappa (2012) chose *L. danicus* and *L. aporus*, which showed an almost contrasting time of occurrence (entire year except summer and autumn/summer respectively) to further investigate their growth and metabolomes under three different temperatures, 12 °C, 19 °C, 26 °C. The temperature range was chosen based on the temperatures recorded in the GoN during the year.

The general outcome was that *L. danicus* appeared to be adapted to grow under a wider temperature range than *L. aporus.* The latter grows equally efficiently at medium and high temperature but less at low temperature, where there was a trade-off between time taken to accumulate the given biomass and biomass yield (more time, higher biomass).

At the same time, contrasting patterns were shown in the metabolomic variations of the two species at different temperatures. Finally, the analysis of the first transcriptomes acquired from one strain of *L. danicus* and one strain of *L. aporus* revealed differentially expressed genes that can be related to the differences already mentioned between the two species, and particularly to life cycle features (Nanjappa et al., in prep.).

1.5. Phenological diversity

Over the year, the abundance of diatom species fluctuates greatly showing peaks (blooms) in certain seasons and sudden drops in others. Even though a general idea of species seasonality might be possible, the exact seasonal succession patterns and bloom events remain largely unpredictable (Rengefors and Anderson, 2006). Each species has traits that allow it to thrive under specific environmental conditions, so the reoccurrence of these environmental conditions over the seasonal time stimulates its growth at specific times (Margalef 1978; Reynolds, 1998; Kneitel and Chase, 2004). Ribera d'Alcalà et al.(2004) suggested that changes in physical factors lead to a change in the water column structure and/or in residence times of water bodies in the Gulf of Naples playing an important role in interannual variations. Despite the quantitative interannual variability, the occurrence of species was regular, indicating that resource availability explains only part of the temporal variance of plankton. Biological rhythms could regulate the temporal dynamics of the communities, whereas the abiotic forcing could modulate the amplitude of the growth phases (Ribera d'Alcalà et al., 2004) In addition, some species show many peaks over the year, even in seasons that differ in terms of environmental parameters and others show long bloom times lasting even months, covering a wide range of environmental conditions (McDonald et al., 2007; Ribera d'Alcalà et al., 2004). In the case of Leptocylindraceae, there is high seasonal diversity with some species being detected almost throughout the whole year, such as *L.*

danicus, and other species showing a more specific seasonal preference, such as *T. belgicus,* or species showing an almost contrasting seasonal preference, such as *L. aporus* and *L. danicus*. Mainly thanks to molecular approaches, many recent studies have revealed cryptic or pseudo-cryptic diversity in diatom species, , which are often reflected in remarkable ecological and physiological differences (Sarno et al., 2005; Amato et al., 2007; Nanjappa et al., 2013; Vanelslander et al., 2009; Degerlund et al., 2012; Huseby et al. 2012; Mann and Vanormelingen, 2013).

Recently, Ruggiero et al. (2015) used a DNA metabarcoding approach coupled with clone libraries to depict diversity and temporal patterns of species of the genus *Pseudo-nitzschia*, a cosmopolitan diatom genus that includes many cryptic and pseudo-cryptic species. High seasonal diversity was detected among species from the GoN, some of which recorded in two or more consecutive periods of the year and others found in single periods. In agreement with results shown in Ribera d'Alcalà et al. (2004) based on LM species detection, the seasonal occurrence of *Pseudo-nitzschia* species appeared not to be constrained by nutrient levels, which is expected particularly for species that contribute little to the total phytoplankton abundance because their regularity strongly contrasts with the hydrographic variability in coastal waters. Temperature and photoperiod, which are linked to astronomical cycles, may be related to species occurrence over the seasons directly or indirectly through endogenous clocks (Anderson and Kaefer, 1987). The alternation of cryptic species or even ribotypes of the same species along seasons might reflect adaptations to particular ecological/seasonal niches (Rynearson et al., 2006; Huseby et al., 2012; Degerlund et al., 2012). Hendry and Day (2005) discuss the idea of isolation and adaptation by time caused by different heritable reproductive times of individuals leading to a reduction of the gene flow (transfer of alleles or genes from one population to another) between early and late reproducers. This isolation would eventually produce populations occurring at different times of the year. Isolation by time could also be an important speciation driver and could explain the striking number of cryptic, closely related lineages in unicellular microalgae. More studies on the eco-physiological characteristics of cryptic species that show differences in their ecological/

seasonal niches are needed in order to better understand the mechanisms and processes involved. The current study aims to aid towards the fulfilment of this purpose.

1.6. Plasticity and adaptation

Adaptation and plasticity are of vital importance to be clarified here since one of the main aims of the present thesis is to analyse the responses of Leptocylindraceae strains to different temperatures and determine whether the capacity to thrive under different temperature conditions in the natural environment depends on phenotypic plasticity or instead that species consist of populations, each with a rather narrow physiological tolerance and response (i.e. possibly adapted populations). Here we define "population" as a group of conspecific individuals that is demographically, genetically, or spatially disjunct from other groups of individuals, "plasticity" as changes in phenotype without any underlying change in genotype and "adaptation" as changes in phenotype which are caused by changes in the genotype.

While plasticity and adaptation are basic terms involved in the context of functional intraspecific diversity, they are also relevant to theories trying to explain the plankton paradox stated by Hutchinson (1961), one of the biggest questions in marine ecology. The paradox describes how a limited range of resources supports an unexpectedly wide range of plankton species defying the competitive exclusion principle on the extinction of one of two species that compete for the same resource. Many microalgal species are differentiated into genetically distinct populations with some of them actually containing both cosmopolitan and endemic clades (Godhe et al., 2006; Rynearson et al., 2009; Watts et al., 2011). The mechanisms behind the high intraspecific genetic diversity are largely unknown (Figuerola and Green, 2002). The assumptions that dispersal rates are high and the extinction of local populations is therefore negligible in the marine environment are controversial and need further investigation. Local adaptation has been found to restrict populations in certain habitats (Foissner et al., 2008; Weisse et al., 2008) and can be either related to geographic distance as in the diatom *Pseudo-nitzschia pungens* (Casteleyn et al., 2010) and the coccolithophore *E. huxleyi* (Iglesias-Rodriguez et al., 2006) or circulation patterns as in the diatom *Ditylum brightwellii* (Rynearson and Armbrust, 2004) and the dinoflagellate *Alexandrium*

fundyense (Richlen et al., 2012). So, as in the monopolization hypothesis of De Meester et al. (2002), which was developed for freshwater zooplankton, gene flow can be largely uncoupled from dispersal in marine phytoplankton since local adaptation and numerical effects of residents may strongly reduce or even prevent successful invasion (immigration). This hypothesis argues that large genetic differentiation even between populations in well-connected habitats is explained by the rapid population growth; fast growing populations that are newly established can adapt quickly to the local environment conditions (Haag et al., 2006).

Yet of course, the maintenance of taxon diversity in sympatric populations cannot be explained by local adaptation. Hutchinson himself suggested that plankton communities could not be in equilibrium due to weather-driven fluctuations. Scheffer et al. (2003) support this notion, suggesting that ecological and environmental factors continually interact such that the planktonic habitat never reaches an equilibrium in which a single species is favored. In practice the homogeneous well-mixed conditions assumed in the competitive exclusion principle hardly exist, since even the open ocean has a spatial complexity facilitating coexistence of species. However, the addition of spatial structure to a simple model of resource competition can provide even an equilibrium solution to the Hutchinson's plankton paradox. Tilman (1977) experimentally tested planktonic algae grown along a two-resource gradient (phosphate and silicate) and suggested that when several species compete for the same resource, the superior competitor at equilibrium should be the species with the lowest requirement for the resource. Several other early studies supported the same model (O'Brien, 1974; Hansen and Hubbell, 1980; Hsu et al., 1977). Tilman et al. (1981) also demonstrated that a species may be a superior nutrient competitor through only a part of the temperature range in which it can survive in the absence of competition, suggesting that "such physiologically caused specialization on a particular nutrient ratio can allow the equilibrium coexistence of many more species than there are limiting resources if a habitat is spatially structured". Additionally, the study of phytoplankton populations in the deep Lake Constance revealed that all factors studied (temperature, stratification, light, competition for resources, sedimentation, losses to heterotrophic compartments of the trophic chain) were

important at certain times of the year for their seasonal succession, but there was a hierarchy among them, while the seasonal pattern of stratification and mixing was the main environmental variable (Sommer, 1985).

Although both nonequilibrium and equilibrium solutions to the paradox indicate that the coexistence of numerous algal species is no longer paradoxical, it is still unknown which of the various equilibrium and nonequilibrium approaches are able to predict the population dynamics, seasonal succession, and other aspects of the structure of natural algal communities (Tilman et al., 1981). The interactions described so far are local rather than global, while all previous game-theoretic models assume that interspecies rankings are fixed and identical across all individuals, excluding individual variability. However, interactions in an ocean do not remain local and competing species without individual variability, which are periodically mixed, have been shown to cease to coexist (Kerr et al., 2002; Menden-Deuer and Rowlett, 2014). Intra-specific variability can be observed as niche differentiation, variability in ecological and environmental factors and variability of individual behaviors or physiology. The outcome of competition among individuals with variable competitive abilities is unpredictable; this unpredictability results in the observed survival of individuals and persistence of diverse species. Therefore plasticity, which results from intra-specific variability, could complete the previously mentioned explanations of the paradox of the plankton since individual variability would maintain high functional diversity within a species making it possible to persist in a habitat despite constant competition with several other planktonic species (Menden-Deuer and Rowlett, 2014). Highly plastic organisms are favoured, especially in fluctuating environments, as they perform well in a variety of habitats because they can rapidly acclimate to new environmental conditions (Pigliucci, 2005; Reusch and Boyd, 2013; Schaum et al., 2013). In order to better explore these issues, it is fundamental to first understand what are plasticity and adaptation and in what way they are linked or diversified.

In contrast to plasticity, adaptation involves a change in mean phenotype of a population due to changes in the genetic composition of that population over time. The definition of adaptation is a central question in evolutionary biology. There have been several propositions, but here

adaptation will be defined as an apomorphic (evolutionarily derived) feature that has evolved in response to natural selection for an apomorphic function (Wanntorp, 1983; Pagel 1994). This definition stresses that a trait can be considered as adapted for a function only when it changes in response to a certain selective agent, to fulfil this certain function. In order to further diversify adaptation from plasticity, the former will be linked with genetic changes occurring when natural selection acts on the genetic variability of a population. On the other hand, phenotypic plasticity is the ability of a single genotype to express a range of phenotypes, discrete or continuous, in different environments and it can work as an adaptive strategy to cope with a range of environments (Stearns, 1989). In this frame, plasticity can be adaptive with respect to a function, and may be altered by natural selection and ultimately become or facilitate adaptation, or it can be non-adaptive. The degree to which plasticity is adaptive or non-adaptive depends on whether environmentally induced phenotypes are close enough or far away, respectively, from a new phenotypic optimum for directional selection to act on (Ghalambor et al., 2007). Adaptive plasticity, which places populations close to an optimum, is the only plasticity that predictably enhances fitness and is most likely to facilitate adaptive evolution on ecological time-scales in new environments. Environments that fluctuate regularly will favour the evolution of adaptive plasticity, (Reusch and Boyd, 2013). The evolution is likely to depend on the relative frequency of different environments and their relative influence on overall fitness (Houston and McNamara, 1992; Kawecki and Stearns, 1993). Non-adaptive plasticity has received relatively little attention but recent results show that it might also be an important mechanism that predicts evolutionary responses to new environments (Ghalambor et al., 2015).

Phenotypic plasticity delays competitive exclusion; each genotype within a species might react differently to different variables - e.g., temperature - but no genotype will be able to “out-compete” all others consistently across all temperatures as their fitness ranking order changes with temperature (Sebens and Thorne, 1985; Gsell et al., 2012). The fact that each extant genotype can show its own interaction pattern with an environmental modulator of genotype fitness such as temperature suggests a more intricate succession pattern of genotypes, constantly

shifting in a temporally variable environment. Over evolutionary time, there will be a stabilization of generalized plasticity within the species yet not of a specific plastic trait; instead Darwinian selection will act on the existing genetic variation for that trait to be genetically accommodated (West-Eberhard, 2003). This theory implies that fluctuating environments increase plasticity over evolutionary time and advantageous plastic traits will become encoded by the genome, if allowed by the existent genetic variation.

However, recent studies have shown a more direct way from environmental perturbation to genetic accommodation. To explain how the exact same genotype may show different reactions depending on the environmental parameters, epigenetic mechanisms have been proposed as a key answer; caste polyphenisms in social insects, seasonal polyphenisms in butterflies and mechanisms of learning and immune system adaptation are well-known examples of phenotypic plasticity that have been linked with epigenetics changes (Fusco and Minelli, 2010, Kucharski et al., 2008, Simola et al., 2013; Lim et al. 2012). Epigenetic marks include DNA methylation, histone modifications and small RNA molecules that lead to reversible modifications on DNA or histones affecting gene expression without altering the actual DNA sequence (Duncan et al., 2014). Epigenetic mechanisms have the potential to define and alter cell phenotypes while the epigenome (the record of epigenetic changes) can be altered by the environment. Therefore through regulation of gene expression they may coordinate a dynamic regulation of the genome in response to environmental changes (Duncan et al., 2014). Epigenetics might be an essential tool that cells use in order to “remember” a past gene-regulatory event and the corresponding stimulus, or hold one in reserve until it is needed (Anway et al., 2005; Stouder and Paoloni-Giacobino, 2010; Manikkam et al., 2012). Epigenetic modifications are inherited by daughter cells during cell division but are usually “reset” at each generation and not transmitted from one generation to the next through sexual reproduction. However, for a wide range of epigenetic mechanisms and organisms there is evidence that some epigenetic changes escape “resetting” (a subset of epigenetic modifications is transmitted through meiosis to subsequent generations), and persist through sexual reproduction (Jablonska and Raz, 2009; Chandler and Stam, 2004;

Vastenhouw et al., 2006; Buckley et al., 2012; Anway et al., 2005; Nelson et al., 2012; Castel and Martienssen, 2013). Epigenetic changes that can be transmitted across generations can also affect adaptation (Bonduriansky, 2012; Jablonska and Raz, 2009).I It has been shown that when natural selection acts both on epigenetic and genetic level, adapted phenotypes show up much earlier than genetic changes do, so populations are able to adapt faster than the cases of selection only on genetic variation (Klironomos et al., 2013). Therefore, epigenetics or otherwise plasticity can affect the timing of adaptive genetic changes and offer responsiveness even in the absence of standing genetic variation or the ability to generate it rapidly. Testing experimentally if and how genetics of adaptation is influenced by epigenetic mutations remains a challenge since long-term evolution is already known to involve genetic change but we have no idea on what proportion of phenotypic adaptation relies immediately on genetic variation.

Having tried to make as clear as possible the relation between adaptation and phenotypic plasticity, the exact use of the terms is defined for the current study as follows:

- Physiologically plastic populations refer to populations consisting of genotypes with high phenotypic plasticity in terms of their physiology.
- Adapted populations are populations that have already reached to the state of acquiring the necessary genetic variations that will help them flourish in a specific environment with possibly no ability to adjust to other environmental parameters that diverge in a high degree from the ones already adapted to.

1.7. New technologies: contribution to the diversity study of *Leptocylindrus* species

Sequencing technologies have brought a revolution in the field of ecology, including marine ecology as well. Marine organisms that are abundant and even critical for our survival are little understood, rarely cultured and described, and sometimes yet to be even seen (Keeling et al., 2014). The use of molecular sequence data can confront these problems (Moustafa et al., 2009).

Although sequencing of the whole genome is possible and can offer an enormous amount of new information on the ecology and evolution of a species (Armbrust et al, 2004), sequencing only a small, standard part of the genome as a marker, called DNA barcoding, is preferential when the main aim is a species-level DNA based identification or the exploration of species diversity . In-depth analyses using molecular markers have offered much information regarding the diversity of marine phytoplankton species; in all cases to date a high degree of interpopulation diversity within dominant phytoplankton groups has been discovered (Rynearson and Armbrust, 2004; Wohlrab et al., 2016).

Following the sequencing of the genome or part of it, the sequencing of the transcriptome (transcriptomics) came to be used in order to better understand the responses of diatoms to nutrients and other biological, chemical and physical variables that characterize the ocean environment (Dyhrman et al., 2012; Lommer et al., 2012). Transcriptomics can also be rephrased as expression profiling since it is referred to the study of the expression level of mRNA transcripts in a given cell population. A first advantage of transcriptomics over genomics is that nuclear genomes are more difficult to be sequenced and assembled and gene modeling is not always straightforward, making transcriptomics the best alternative way to generate a marine microbial reference database of expressed genes coming from pure cultures, of which we are in a great need (Keeling et al., 2014). Transcriptome sequencing not only offers a comprehensive analysis of differential gene expression among conditions or species, but also promises for the annotation and quantification of all genes along with their alternative isoforms across samples and for the discovery of novel genes, since transcriptome data do not contain introns (Garber et al., 2011; Kim et al., 2014). The Joint Genome Institute (JGI), supported by the United States Department of Energy, provides transcriptome data for many algal species, whereas the National Center for Genome Resources (NCGR) and the Gordon and Betty Moore Foundation's (GBMF) Marine Microbiology Initiative (MMI) hold a transcriptome sequencing program from 750 marine microbial eukaryotes including many algal species (http://www.marinemicroeukaryotes.org). The data in these databases are freely available.

Since transcriptomics and in particular RNA-sequencing technique are quite recent developments, improvements in the generation, manipulation and analysis of the data are still ongoing, resulting in a great variability in the maturity of the available computational tools (Garber et al., 2011). Therefore, different methodologies can have an impact on the results and interpretation of the data. Nevertheless, transcriptomics is being largely utilized in marine plankton research. Several diatom based studies taking advantage of transcriptomics have been conducted the last decade, mainly focusing on changing environmental conditions such as nitrogen or iron starvation (Mock et al., 2008; von Dassow et al., 2009; Bhadury et al., 2011; Lommer et al., 2012; Bender et al., 2014; Buhmann et al., 2014; Di Dato et al., 2015; Levitan et al., 2015; Wohlrab et al., 2016). It seems indeed that studies based on data produced by sequencing technologies will continue for years to come, considering that the sequencing detection and resolution power is constantly improved while the related problems and limitations are better understood and confronted.

A rapid, and therefore widely used, method of biodiversity assessment is DNA metabarcoding, which takes advantage of DNA barcoding using universal PCR primers to amplify and identify DNA from mass collections of organisms or from environmental DNA. Sequencing a whole community coming from an environmental sample, at a single point in time and space focuses on a higher level and allows investigation of organisms while in their natural condition and environment; metabarcoding provides an indication of who is there and how diverse they are, overcoming the bias against rare taxa. Subsequently, the sequencing of the transcriptome of a community (metatranscriptomics) would ideally tell us how the species behave in their environment without any effect by the culturing conditions. In the end, the exploitation of these technologies should provide a better understanding of the whole diatom or even the whole plankton community, including interactions between species and with environment. Another aspect that could be addressed by these approaches is the clarification of the importance of the rare biosphere. Communities can be characterized by few dominant taxa, and many low-abundance but highly-diverse taxa, a group which is referred to as the “rare biosphere” (Pedròs-Aliò, 2007; Sogin et al., 2006), which can make the difference since it might be equally or even more active in terms of

key physiological functions in the environment than the abundant species. In addition, metagenomics studies have led to the seed bank hypothesis which suggests that environments do not contain only species adapted to them but also species that reached the area and did not flourish but survived. The seed bank consists of microbes that can enter dormancy or low metabolic activity stages and eventually bloom when the environment changes. Caporaso et al. (2012) proved that the differences in the community composition can be due to changes in the relative abundance of taxa that there were always present in the environment. Despite the remarkable scope of metabarcoding, there are still plenty limitations for this approach such as the substitution of the concept of species with OTUs (Operational Taxonomic Unit, a term used for groups of closely related individuals) and their reliability, the dependence of results on the DNA extraction protocol, the low depth sequencing compared to the quantity of DNA in a sample, the lack of annotation data and functional verification for sequence annotation (Zarraonaindia et al., 2013; Delmont et al., 2012; Warnecke and Hugenholtz, 2007). A complementary approach to the taxonomical determination of natural communities will be to detect environmentally restricted functional genes and pathways in order to detect metabolic activities in distinct environments (Dinsdale et al., 2008; Rusch et al., 2007). To sum up, as it is already stated in Zarraonaindia et al. (2013)'s review article, individual omic techniques are not able to clarify the complex aspects of microbial ecosystems so a multilevel omic approach is needed together with a spatiotemporal sampling design. The present study takes advantage of the two omic techniques, transcriptomics and DNA High Throughput Sequencing (HTS) metabarcoding, and although they are not directly combined, an attempt to combine the information provided by them on functional and genetic diversity, respectively, was done with the final goal of capturing the general diversity (in all levels) and thus better understanding the ecology and evolution of the study species. Molecular knowledge and technologies have been exploited with notable results so far, helping to a great extent to unfold the genetic and functional diversity of the Leptocylindraceae species. Therefore, molecular technologies were used again and constituted the main tool of the continuation of the Leptocylindraceae universe exploration.

1.8. Aim

This research project aimed at exploring the physiological, functional and molecular diversity of the Leptocylindraceae species and thus, identifying the related differences across the seasonal and geographical variations of environmental parameters. The reasoning behind this approach is that similar or distinct diversity patterns within and among species would shape, or be a result of, their composition (structure) across the seasons and the different places. To this end, each chapter dealt with one or more of the mentioned levels of diversity (**physiological, functional, molecular**) in the following way:

- Chapter 2 investigated **physiological** diversity of two of the most abundant Leptocylindraceae species, *L. aporus* and *L. danicus,* through growth experiments performed at different temperatures, including low and high temperature. The two species are known to show a broad but at the same time a contrasting time of occurrence in the Gulf of Naples (GoN). Therefore, the results of this chapter could explain in what way a species manages to maintain intermediate to high levels of abundance through the constantly changing environmental conditions. For each species, multiple strains isolated in different seasons in GoN were used in order to understand if the two dominant species could be composed by strains adapted to the environmental temperature of each season or rather by highly plastic strains that could respond similarly or differently to each other. The last case of different responses would indicate intraspecific physiological diversity, independent of the seasonal origin of the strains and of any possible genetically-based adaptation.
- Chapter 3 explored the **functional** diversity of three *L. aporus* strains used in the growth experiments of Chapter 2 by obtaining the transcriptome of each strain at the three different temperatures and then performing a differential expression analysis between temperatures and between strains as well. In that frame, the results could associate with the ones of the previous chapter. This analysis would reveal the specific responses at different environmental conditions of an abundant species with a broad temporal range

in GoN. Certain functions or mechanisms and pathways could be identified and linked to heat or cold reaction and even adaptation, which could help us understand the strategy followed by such species. Furthermore, the between-strains analysis could show a similar behaviour for all strains or strain-specific gene expression profiles, revealing the level of intraspecific functional diversity in *L. aporus* and its possible role in the ecology of this species.

- Chapter 4 examined the intra- and interspecific **functional** and **molecular** diversity of four *Leptocylindraceae* species, *L. danicus, L. aporus, L. convexus* and *L. hargravesii.* All strains used were isolated in GoN and grown under the same conditions. Comparative transcriptomics was used in order to explore the functions within and among species while phylogenomics could show any molecular diversity based on microvariations. *Leptocylindrus convexus* and *L. hargravesii* are more scarce in GoN while they have a more restricted seasonal niche compared to the other two species. Therefore, the analysis of this chapter could reveal functional traits and diversity patterns related to the specific characteristics of each species while it could also uncover similar or different strategies among them.
- Chapter 5 dealt with the **molecular** diversity of all Leptocylindraceae species and their corresponding temporal and spatial distribution. DNA HTS metabarcoding was used to explore for Leptocylindraceae in data gathered at Tara and BioMarKs world stations and several dates during three years at the long-term ecological research MareChiara (LTER-MC) station in GoN. This analysis could lead to the identification of new ribotypes or even new Leptocylindraceae species, which would broaden our knowledge of intra and interspecific genetic diversity. The geographic and seasonal distribution of each ribotype or species would represent the ecological aspect of the genetic, phenotypic and functional diversity described in the previous chapters and therefore, help determine any signs of ecological adaptation to the different environmental conditions found throughout the year and/or at different localities.

In the end, the knowledge and information gained by the analyses carried out in the chapters described above should provide some answers to two important questions:

a) *Are individual species able to respond to environmental fluctuations due to physiological plasticity or genetically-based adaptation to different conditions?*

Individual species have been found to show different but also recurrent distribution patterns through the year while at the same time show specific spatial distribution across the world. Given the fact that plasticity or adaptation would be reflected on the homogeneous or specific, respectively, response of the Leptocylindraceae species to environmental changes, Chapter 2 and Chapter 3 would give a first insight into the answer to this question. Any intraspecific physiological or functional diversity detected through these analyses could serve as a link either to plasticity or to adaptation in the case of a correlation with strains coming from a certain period of the year. Furthermore, the spatial and temporal distributions defined in Chapter 5 would express the ecological side of the diversity produced by plasticity or adaptation.

b) *Are the different co-occurring species reacting in the same way under the same environmental conditions or each one of them takes advantage of different cues of the environmental spectrum and therefore have different functional patterns?*

At the station LTER-MC, diatom blooms consist of several species each time. So the answer to this question could come from Chapter 4 where differences in gene content and/or gene expression between the Leptocylindraceae species might be linked to their different physiology, seasonality or life cycle. The level of intraspecific functional variability, as well as of the genetic microvariations indicated by the transcriptomes, could be also a characteristic borne by species of a certain seasonal occurrence and spatial distribution.

Summing up, the present study would reveal the presence of high or low diversity in several levels, both within and among Leptocylindraceae species, which could serve as evidence for its most probable source, either that being plasticity or adaptation, the possible relation or

interaction between those levels of diversity, and, most importantly, their connection to the ecology of each species.

Chapter 2. Growth response of *L. aporus* and *L. danicus*

2.1. Introduction

One of the most important factors of growth is temperature. There is a relatively limited range within which species can grow; in some groups like prochlorophyte temperature plays a key role while in others plays a secondary one and nutrients or light seem to be either antagonistic or synergistic factors (Cavender-Bares et al., 2001; Rose et al., 2009). Eppley (1972) and Goldman and Ryther (1976) concluded that temperature has little effect on the production of phytoplankton in the sea while Karentz and Smayda (1984) accept that the seasonal appearance of species is a prime example of how temperature does indeed exert a huge influence on species competition. It has been shown though that the optimum temperature for growth in laboratory cultures can be 3 to 14 °C higher than the field temperature in which the species flourish (Karentz and Smayda, 1984). In addition, Karentz and Smayda (1984) observed that temperature during the annual field maxima varied significantly for all the studied species and concluded that the considerable interannual variation is driven by other controlling factors which also change interannually together with the temperature.

Either way, it is generally accepted that temperature mainly affects cellular structural components, especially lipids and proteins, and reaction rates. Changes at these levels have secondary effects on metabolic regulatory mechanisms, cell permeability, cell composition and specificity of enzyme reactions (Richmond, 1986). An increase in temperature leads to an increase in enzyme activity in metabolic processes, including photosynthesis and respiration so the cells are expected to grow faster (Falkowski et al., 1997). Low temperatures lead to the exact opposite direction so in order to survive, cells set off an increase in enzyme synthesis (DeNicola, 1996). Toseland et al. (2013) used a set of interdisciplinary approaches combining metatranscriptomes with biochemical data, cellular physiology and emergent phytoplankton growth strategies in a global ecosystems model in order to prove that eukaryotic phytoplankton metabolism is highly influenced by temperature. It was concluded that under high temperatures the number of ribosomes and their associated rRNAs decreases but the rate of protein synthesis increases, whereas under low temperatures the mRNA translation efficiency is reduced but it is partially

compensated by an increase in cellular concentrations of ribosomal proteins. However, too high temperature can also lead to reduction of algal growth; this reduction is related to denaturation and degradation of certain proteins (Downs et al., 2013), reduced functionality of the photosynthetic machinery (Rowland et al., 2010), decreased RUBISCO activity, stimulated respiration (Davison, 1991) and disturbed functions of cell membranes (Glatz et al., 1999).

In diatoms, the effect of temperature is still not clear for most species. The maximum growth rate change can be considered a species-specific characteristic; cell size might be a determining variant with smaller diatoms having higher growth rates due to their catalytic advantage (Sarthou et al., 2005). The upper temperature limit is different for each species but culture experiments have already proven that clones of the same species isolated from warm or cold waters may also differ in their optimum temperature for growth (Braarud, 1961; Hulburt and Guillard, 1968). The interactions and exact relationship between temperature and other environmental parameters such as nutrients and light are not yet well characterized despite the many experiments showing the significant effects (synergistic or antagonistic) on growth, photo-physiology and calcification (Fu et al., 2007; Hare et al.,2007; Rose et al.,2009; Feng et al.,2009). Diatom distribution is influenced by all these environmental parameters and even though the role of temperature might differ compared to the rest for each species, it is definitely a factor that either directly or indirectly has an effect on their growth and physiology.

In this Chapter, growth experiments on *L. aporus* and *L. danicus* strains collected at different periods of the year were performed at 13 °C, 19 °C and 26 °C. The results gave a first idea of the strains' responses to different temperatures and whether they respond differently due to their physiological plasticity or rather they are different populations adapted to different conditions (see Chapter 1 for further information on plasticity and adaptation).

2.2. Materials and Methods

2.2.1. Isolation and molecular characterization of strains

Before starting with the experiments, establishment of *Leptocylindrus* species cultures from coastal waters of the Gulf of Naples was performed in the first months of the winter (December 2013-February 2014) with the aim to expand the SZN collection and isolate new strains of *L. aporus* and *L. danicus* for the growth experiments and also of *L. hargravesii,* which was at that time missing from the collection. Phytoplankton is regularly sampled (bottle and net samples) at LTER-MC (MareChiara) station in the GoN on a weekly base. From net samples, which contain mainly >20 µm fraction of the phytoplankton assemblage, cells recognized in the light microscope as *Leptocylindrus* were picked up by a capillary tube, placed into culture medium droplet and transferred several times into new separate droplets in order to eliminate any contaminant species. When a single cell was isolated, *in vitro* cultures were attempted to be established by progressively increasing the volume of the medium (first grown in a 12-well culture plate, then 6-well and finally in 25 cm^2 flasks).

The medium used for the maintenance of the cultures was initially F/2 medium (Guillard and Ryther, 1962) prepared on sterile water plus silica, but was later replaced by K medium (Keller et al., 1987) plus silica, based on the observations by the laboratory staff of a better growth of *Leptocylindrus* species in this specific medium. The cultures were kept at 20 °C, under fluorescent light of 100 µmol photons m^{-2} sec^{-1} and a photoperiod of 12:12 (light: dark) and were refreshed regularly (weekly or every second week).

Total DNA was extracted from individual strain cultures following the C-TAB DNA extraction protocol:

Materials

- C-TAB extraction buffer (2% CTAB, 200 mM Tris HCL pH 8.0, 50 mM EDTA, 1.4 M NaCl, 2.5 PVP)
- β- mercaptoethanol
- Chloroform: isoamyl alcohol (SEVAG), 24:1

- Isopropanol (-20 °C)
- 75% ethanol (-20 °C)
- dd H_2O (50 °C)

Protocol

1. 1-1.5 ml of cell culture were collected and centrifuged for 15 minutes at 5000 rpm (2600 x g). Supernatant was discarded.
2. The pellet was resuspended in 500 ml C-TAB and 12 µl β-mercaptoethanol and vortexed (under hood).
3. Incubation was performed at 65 °C for 35 min. The material was vortexed in between and at the end of incubation.
4. 500 µl of SEVAG was added and tubes were shaken gently. Centrifugation followed at 14000 rpm (20200 x g) for 20 min. The upper layer was pipetted out and transferred into a new tube.
5. 400 µl of cold isopropanol was added and then mixed gently. Centrifugation followed at 14000 rpm (20200 x g) for 20 min. Supernatant was discarded.
6. The pellet was washed with 400 µl of 75% cold ethanol and centrifuged again at 14000 rpm (20200 x g) for 15 min. Alcohol was removed and sample was air-dried for 30-40 min (under hood).
7. 40 µl of dd H_2O (first warmed at 50 °C) were added and samples were finally stored at -20^{o}C.

Following DNA extraction, the Internal Transcribed Spacer region (ITS) of the nuclear ribosomal RNA-coding cistron was amplified in order to identify the species based on this molecular marker. This specific marker was selected because, according to Nanjappa (2013), it is the most variable one and also allows distinguishing between *L. danicus* and *L. hargravesii*, which show base changes at least at 29 positions in the ITS. The specific primers and PCR protocol follow:

<table>
<tr><th colspan="4">Primers</th></tr>
<tr><td colspan="2">LDITSF (Nanjappa, personal communication)</td><td colspan="2">5’-ATTACGTCCCTGCCCTTTGT-3’</td></tr>
<tr><td colspan="2">ITS4 (White et al., 1990)</td><td colspan="2">5’-TCCTCCGCTTATTGATATGC-3’</td></tr>
<tr><th colspan="4">Amplification Protocol</th></tr>
<tr><th colspan="2">Reagents</th><th colspan="2">PCR settings</th></tr>
<tr><td>PCR buffer (10x)</td><td>2.5 µl</td><td>94°C</td><td>5 min</td></tr>
<tr><td>dNTP (10x)</td><td>2 µl</td><td>94°C</td><td>30 sec</td></tr>
<tr><td>LDITSF (10 µM)</td><td>0.5 µl</td><td>44°C</td><td>30 sec</td></tr>
<tr><td>ITS4 (10 µM)</td><td>0.5 µl</td><td>72°C</td><td>2 min</td></tr>
<tr><td>Taq Polymerase (5u/µl)</td><td>0.5 µl</td><td>72°C</td><td>10 min</td></tr>
<tr><td>H_2O</td><td>18 µl</td><td colspan="2" rowspan="2">step 2- step 4: repeated for 40 cycles</td></tr>
<tr><td>DNA (10ng/µl)</td><td>1 µl</td></tr>
</table>

The PCR product was run on a 1% agarose gel. The band was then cut with a razor and DNA was extracted using DNA Isolation Spin-Kit Agarose by AppliChem. Finally, the DNA was quantified on a NanoDrop Spectrophotometer (Thermo Fisher Scientific Inc, UK), then prepared (7 µl DNA of at least 15 fM + 3 µl primer, 10 µM) and sent for sequencing. The sequences received were edited using Bioedit v.7.0.2 and blasted in National Center for Biotechnology Information (NCBI) database.

2.2.2. Growth experiments

For the growth experiment of *L. aporus* six strains isolated from summer and winter samples were used and duplicates were kept for all of them during the acclimatization. For *L. danicus* three strains were selected, one from summer and two from winter samples. Strains were acclimated to three temperatures, 13 °C, 19 °C and 26 °C, at a light intensity of 100 µmol photons m^{-2} sec^{-1} and a photoperiod of L:D, 12:12 (Fig. 2.2.2.1). The temperatures were chosen to test the growth of the strains under natural conditions, i.e. within the range they may experience in the GoN, with no intention to test stressful conditions. As cultures were routinely maintained at 20 °C, before reaching the two final extreme temperatures, cultures spent one week at intermediate temperatures, namely 15 °C and 23 °C. To this aim, cultures were kept in 20 ml of K + Si medium

in glass tubes and waterbaths were used for regulation of temperature. Chlorophyll fluorescence was measured daily using a Turner 10-005 fluorometer.

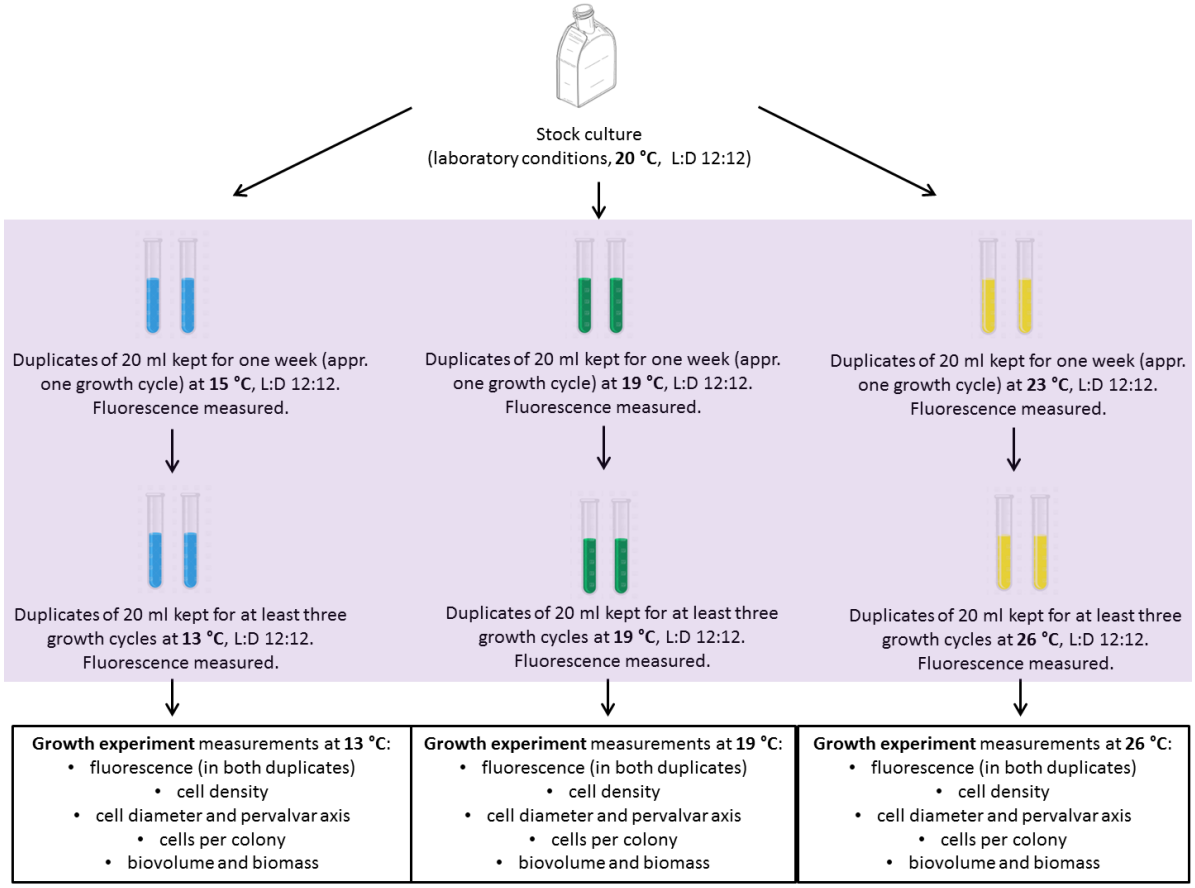


Figure 2.2.2.1 Diagram of the acclimatization (purple shaded area) and growth experiment including all measurements taken for one of the duplicates of each strain.

Cultures were transferred to fresh medium when they were at their exponential phase and just before the stationary phase. The time at which the latter phase would start was estimated based on previous observations of the growth of each strain in each temperature. Growth rate was calculated based on plots of the logarithmic values (base 10) of fluorescence along the incubation time. The slope calculated from the linear trendline applied to this graph is equal to K10. K10 was used in the following equations to calculate growth rate (Andersen et al., 2005):

$k\ (\text{div./day}) = 3.322 / K10$ (1)

and

$Ke\ (\text{day-1}) = 0.6931 * k$ (2)

After the growth rate had remained steady for three growth curves the acclimatization was considered to be completed and the actual experiment started. During the experiment,

fluorescence was measured daily and cell density was calculated by counting on a Sedgewick rafter counting cell slide every other day. Growth rate was calculated as described for the acclimatization, once based on fluorescence values and once based on cell counts. While counting pictures were taken and a record of number of cells per colony, cell diameter and pervalvar axis for 25 randomly selected cells was kept. Mean cell diameter and pervalvar axis were used for the calculation of the biovolume according to the formula for cylinders, V=π*r2*h (Hillebrand et al., 1999). Biomass production is directly dependent on the physiological status of the cell and is thus influenced by abiotic factors. Therefore, the biomass achieved under different conditions was calculated as Carbon converting biovolume using (Menden-Deuer and Lessard 2000) formula; log10C = -0.541 + 0.811 x log10V. The converted log10C gives pg of C cell-1, which multiplied by maximum cell density ml-1 gives the maximum biomass attained in terms of pg of C ml-1.

2.3. Results

From the end of December 2013 until the end of February 2014, 65 strains of *Leptocylindrus* species were isolated and successfully brought into culture conditions. After being characterized molecularly, 16 *L. danicus* (24.6%), 17 *L. aporus* (26.2%), 9 *L. hargravesii* (13.8%) and 5 *L. convexus* (7.7%) were identified in the unialgal cultures established from plankton samples. The high number of *L. aporus* isolated in December 2013 – February 2014 was a surprise since it is considered a summer-autumn species based on previous isolation dates of the species (Nanjappa et al, 2013). 18 strains (27.7%) remained unidentified due to contamination or bad quality DNA/ sequences. No *Tenuicylindrus belgicus* strain was isolated or identified.

The *L. aporus* strains selected for the growth experiment were the following:

1. B651, isolated on 21/8/2010.
2. B704, isolated in October of 2009.
3. B764, isolated on 18/11/2010.
4. 1A1, isolated on 20/12/2013.
5. 3A6, isolated on 28/01/2014

6. 1089-10, isolated on 14/01/2014.

The *L. danicus* strains selected for the growth experiment were the following:

B650, isolated in 15/06/2010

4B6, isolated in 13/02/2014

1089-17, isolated in 14/01/2014

Two of the *L. aporus* strains, B651 and B764, which were maintained in the SZN collection for long time (since 2010) seemed to need more time to acclimatize to the temperature of 26 °C compared to the rest of the strains and needed one more growth cycle. Finally, the growth rate for all strains remained consistent for at least three sequential growth curves (Table 2.3.1). Despite the fact that some growth rates were still significantly different in the last two growth cycles, these strains were considered acclimatized because of a much smaller deviation compared to previous measurements (Table 2.3.1).

Table 2.3.1 Average between duplicates of growth rates, k (div./day), for *L. aporus* strains at the three different growth temperatures inferred from fluorescence values during acclimatization. The strains with the significantly different (Welch and Brown-Forsythe ANOVA, p<0.05) growth rates in the last two growth rates are written in red.

Strain	1st Growth Curve	2nd Growth Curve	3rd Growth Curve
B651 (13°C)	0.42	0.45	0.40
B651 (19 °C)	0.83	0.92	0.86
B651 (26 °C)	0.94	1.06	0.85
B704 (13 °C)	0.38	0.38	0.45
B704 (19 °C)	0.94	1.07	0.98
B704 (26 °C)	0.84	1.15	1.45
B764 (13 °C)	0.36	0.37	0.52
B764 (19 °C)	0.91	1.12	1.06
B764 (26 °C)	0.83	1.04	0.91
1A1 (13 °C)	0.42	0.50	0.50
1A1 (19 °C)	1.05	0.94	0.90
1A1 (26 °C)	0.98	1.11	0.97
1089-10 (13 °C)	0.49	0.51	0.53
1089-10 (19 °C)	1.04	1.23	1.14
1089-10 (26 °C)	1.11	1.28	1.23
3A6 (13 °C)	0.50	0.46	0.50
3A6 (19 °C)	1.21	1.10	0.98
3A6 (26 °C)	1.00	1.02	0.96

The growth curves during the growth experiment for all strains showed a much longer growth cycle at 13 °C compared to the other temperature conditions, while at the same time there was a high intraspecific variability in growth and duration of the cycle under all temperatures conditions (Fig. 2.3.1).

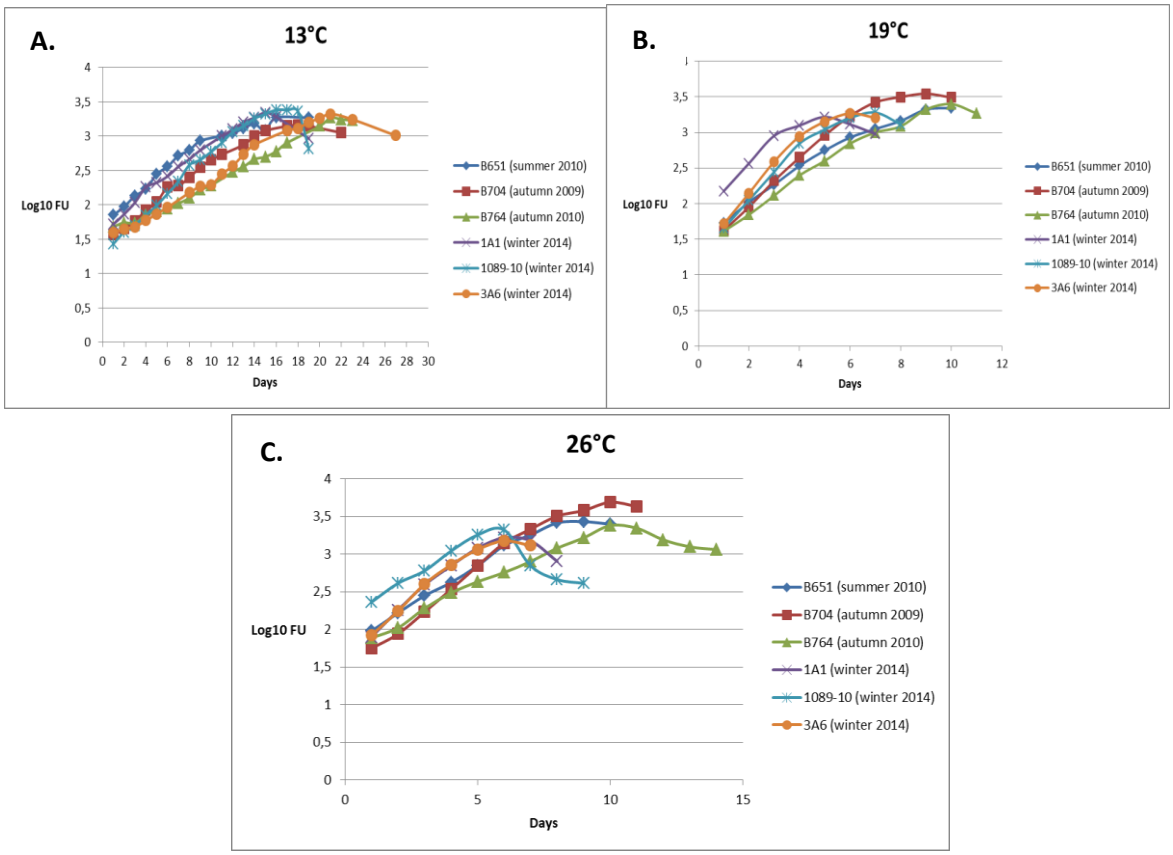


Figure 2.3.1 *Leptocylindrus aporus* strains growth curves established by daily arbitrary fluorescence measurements at three different temperatures, 13 °C (A), 19 °C (B), 26 °C (C). In the legend, the season of isolation of each strain is indicated in brackets.

There were several cases where the growth rate during the experiment deviated considerably from the one observed during the acclimatization period, as shown by the high standard deviations of the growth averaged for B651 at 19 °C and 26 °C, B704 at 26 °C, B764 at 19 °C and 26 °C, 1089-10 at 26 °C and 3A6 at 13 °C (Fig. 2.3.2). The growth rate was significantly different between low temperature and medium/ high temperature (Welch ANOVA, Games Howell post hoc test; $p<0.05$) within and among all strains.

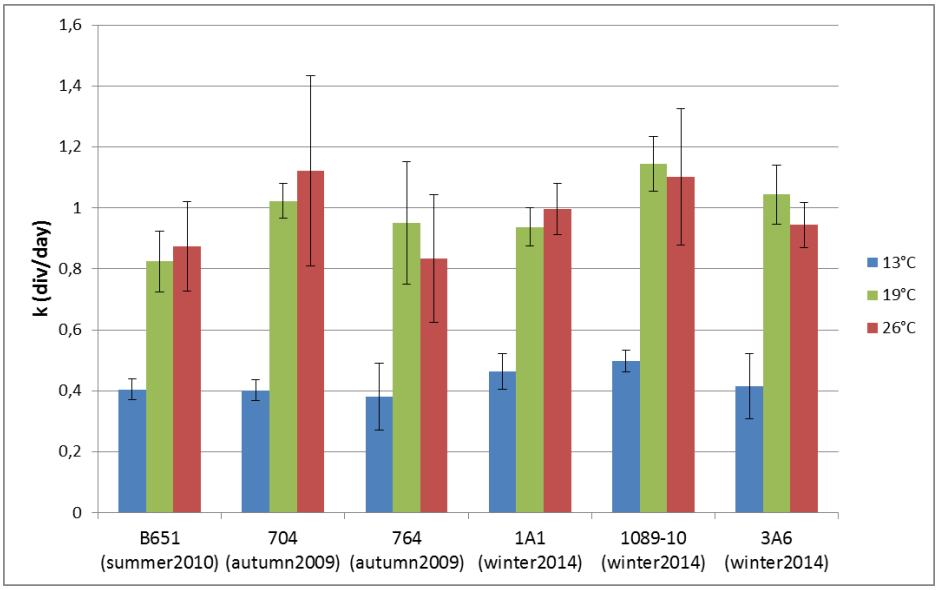


Figure 2.3.2 Bar graph of growth rates of *L. aporus* strains at three different growth temperatures, based on counts. Standard deviation values were calculated based on growth rate values of the four last growth curves (the last three cycles of acclimatization plus the growth experiment cycle). In the legend to the x axis, the season of isolation of each strain is indicated in brackets.

All strains showed a significantly longer exponential phase at 13 °C compared to 19 °C and 26 °C (ANOVA, $p<0.05$) (Fig.2.3.3).

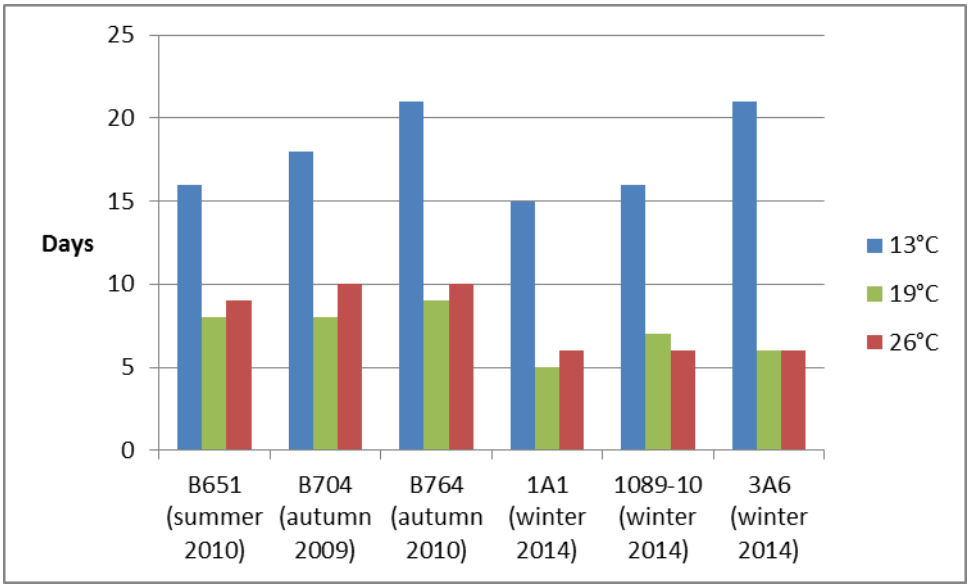


Figure 2.3.3 Duration of the exponential phase for *L. aporus* strains at three different growth temperatures. In the legend to the x axis, the season of isolation of each strain is indicated in brackets.

High differences among the cell density and biovolume values reached at the end of the exponential phase were recorded among different strains and temperatures (Fig. 2.3.4). However, differences in biovolume were not significant among temperature conditions, nor season of isolation (Kruskal Wallis, $p=0.165$).

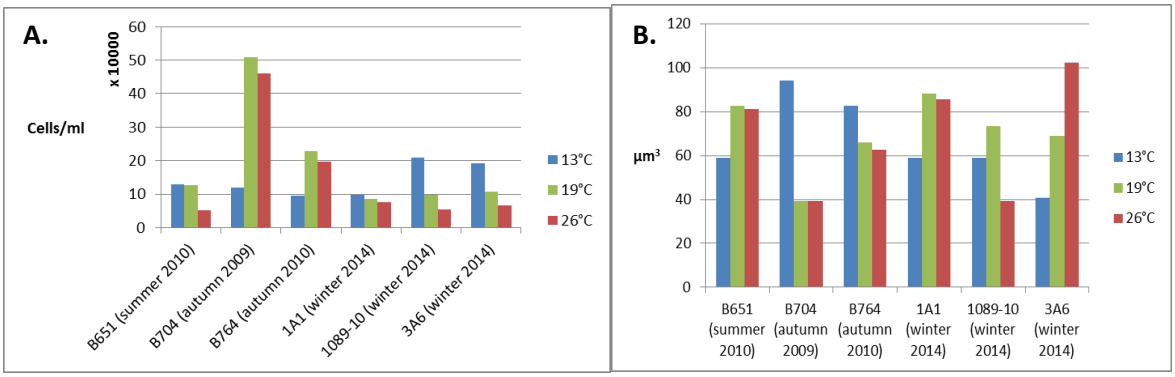


Figure 2.3.4 Cell density (A) and biovolume of cells (B) at the end of the exponential phase for *L. aporus* strains at three different growth temperatures. In the legend to the x axis, the season of isolation of each strain is indicated in brackets.

The biomass yield appeared also to be quite diverse among strains (Fig. 2.3.5). The cell density and biomass were significantly different for autumn strains at 19 °C and 26 °C compared to the rest at 26 °C while the winter and autumn strains were also significantly different at 19 °C; the 13 °C were significantly different to the 26 °C in the winter strains (Kruskal Wallis, $p<0.05$).

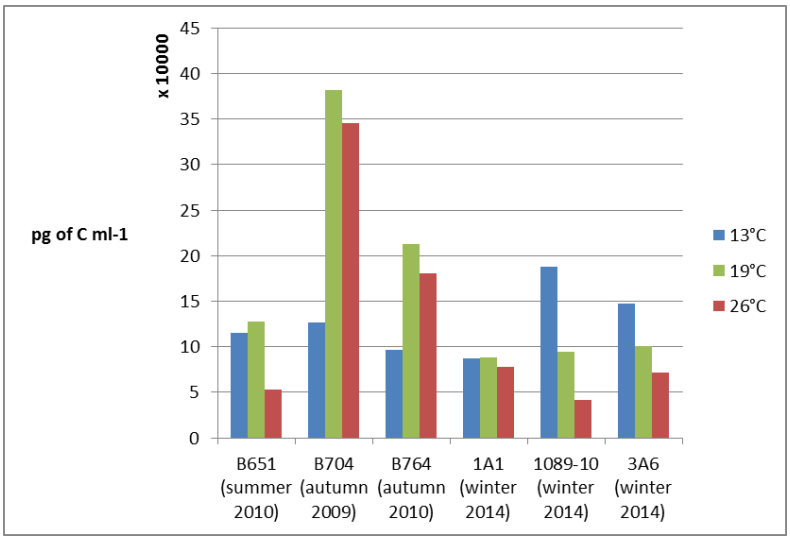


Figure 2.3.5 Biomass produced in the end of the exponential phase of the *L. aporus* strains at the three different growth temperatures. In the legend to the x axis, the season of isolation of each strain is indicated in brackets.

Regarding the cell size and morphology, all strains appeared to form longer colonies and the percentage of larger cells increased at medium and high temperature, with the recently isolated strains forming longer chains (Table 2.3.2). No significant differences were detected between strains isolated in different seasons (ANOVA tests, $p>0.05$).

Table 2.3.2 Size (diameter and pervalvar distance) for 25 randomly selected cells and colony (chain length) condition for *L. aporus* strains at the three different growth temperatures.

Strain	T (°C)	Mean Diameter (µm)	Pervalvar axis (µm)	% of cells with diameter >= 8 µm	Max. diam. (µm)	Min. and Max. Chain Length
B651 (summer 2010)	**13**	5	15	0	5	Single and up to 5 cells
	19	7	15	0	7	Single and up to 10 cells
	26	5	15	37,5	10	Single and up to 12 cells
B704 (autumn 2009)	**13**	8	15	100	8	Single and up to 6 cells
	19	5	10	0	5	Single and up to 7 cells
	26	5	10	0	5	Single and up to 11 cells
B764 (autumn 2010)	**13**	7	15	0	7	Single and up to 6 cells
	19	5	15	20	8	Single and up to 11 cells
	26	5	15	10	8	Single and up to 16 cells
1A1 (winter 2014)	**13**	7,5	15	20	10	Single and up to 22 cells
	19	7,5	15	50	10	Single and up to 25 cells
	26	7,5	15	45	10	Couplets and up to 24 cells
1089-10 (winter 2014)	**13**	5	15	0	5	Couplets and up to 8 cells
	19	5	15	35	20	Single and up to 20cells
	26	5	10	0	5	Couplets and up to 11 cells
3A6 (winter 2014)	**13**	4	13	0	4	Single and up to 11 cells
	19	5	15	17	10	Single and up to 11 cells
	26	10	13	5	20	Triplets and up to 54 cells

For *L. danicus,* the acclimatization experiments ended after 6 weeks due to the strains' incapability to steadily grow under the two extreme experimental conditions. More difficulties were met at the high temperature condition. Therefore, the growth experiment for this species was not completed and only the growth rates during the acclimatization period were recorded (Fig.2.3.6). During this period:

- B650 and 4B6 did not show any significant difference between 13 °C and 26 °C in contrast to 1089-17 (ANOVA, p= p<0.05)
- 1089-17 was significantly different from B650 and 4B6 at 19 °C and 26 °C (ANOVA, p<0.05)
- 4B6 at 19°C was significantly different than 4B6 at 26°C (ANOVA, p=0.014).

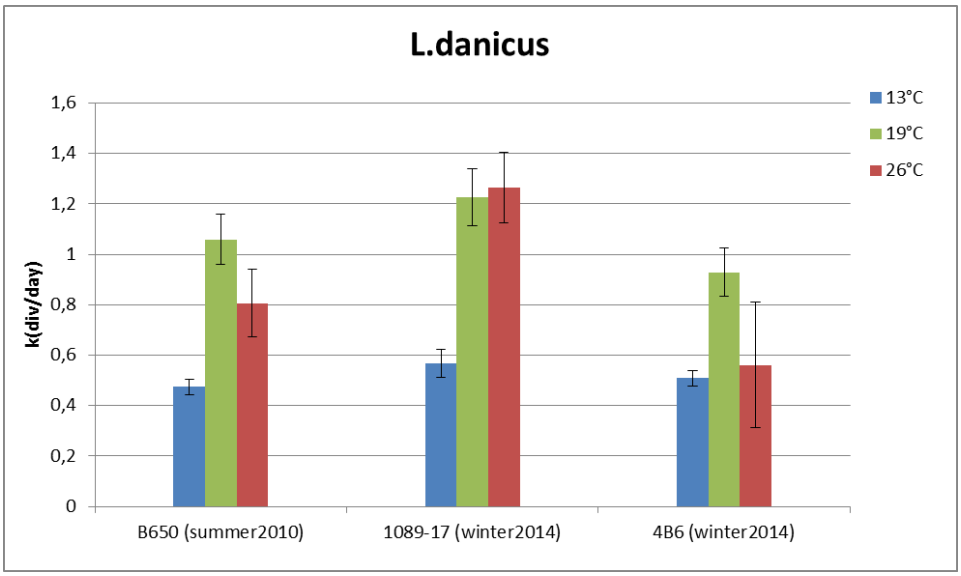


Figure 2.3.6 Bar graph of growth rates of *L. danicus* strains at three different growth temperatures, based on cell counts. Standard deviation values were calculated based on growth rate values of three last growth curves. In the legend to the x axis, the season of isolation of each strain is indicated in brackets.

The mean pervalvar axis and diameter are available only for the 19 °C and they were within the expected range, significantly longer than *L. aporus* (Table 2.3.3).

Table 2.3.3 Size (diameter and pervalvar distance) for 25 randomly selected cells and colony (chain length) condition for *L. danicus* strains at 19 °C.

Strain	Mean Diameter (µm)	Pervalvar axis (µm)
B650	8	28,83
1089-17	8	26,36
4B6	10	29,78

The behavior of the strains was also different and in particular 1089-17 was quite active sexually compared to the other two strains, with cases of very high percentage of spores (almost 80%) present in the culture.

2.4. Discussion

The overall effects of biotic and abiotic factors on cell physiology are reflected in the growth rate and final biomass yield. These are parameters that have been used as metrics for decades to understand the effect of the environment in *in vitro* experiments. Use of one to the exclusion of

the other could lead to misinterpretation regarding the physiological response of the species. *Leptocylindrus danicus* was already known to be adapted to grow under a wider temperature range than *L. aporus,* which is equally efficient at higher temperatures but less at low temperature conditions (Nanjappa, 2012). Starting with *L. danicus* acclimatization period in the current experiment, all strains were affected by the low temperature but two of them were highly stressed also at 26 °C. There was a high variability among them too; 1089-17 was the one growing the best at all temperatures while 4B6 was the one mostly stressed at the high temperature, being the only strain with significantly lower growth rate at 26 °C compared to 19 °C. The different behaviors of the strains including the higher sexual activity of 1089-17 compared to the other two imply that the three strains of the same species were quite different regarding their physiology, responses and ability to adapt/acclimatize. This differentiation does not seem to be linked with the season of their isolation since 1089-17, which was a winter strain, had the better performance at the higher temperature. All *L. danicus* strains were growing slower at 13 °C; a decrease in temperature leads to a decrease in enzyme activity in processes such as photosynthesis and respiration. No other conclusion can be drawn for the interaction of the species with temperature because the strains were still under the acclimatization stress when the measurements were done but the general result was a strain specific behavior in terms of response to temperature.

Before getting into details about the *L. aporus* results it must be noted that the deviation between the growth rate during the experiment and the growth rate observed during the acclimatization period could mean that either these strains were not actually acclimatized or that there was some malfunction in the experimental equipment. In the following discussion we will focus on the results during the experiment leaving aside for now the acclimatization values.

At a first glance, intraspecific variability in the species reactions to the different temperatures is evident. The reason for that will be discussed here, starting from the potential relation to the isolation season of each strain. There was no significant difference between growth rates of different season strains but also no clear correlation to temperature for each season either. The growth rate showed a significant decrease as the temperature decreased but at medium and

higher temperature the values were very close to each other. Indeed for species with a temperature coefficient (Q_{10}, the factor by which a biological rate is increased following a 10 °C rise in the temperature) of about 2, such as *L. danicus* and *L. aporus*, the growth rate for suboptimal temperatures is expected to be slightly lower than the maximum growth rate (Eppley, 1972; Nanjappa, 2012). It seems that both species reached close to their maximum growth rate at a temperature very close to 19°C and possibly remained at this plateau with only slight changes until 26 °C. Despite this remarkable stability of growth rate between 19 °C and 26 °C, there have been indications of different concentration of metabolites in the two temperatures for both species (Nanjappa et al., in preparation). This might imply indeed a change in the enzyme usage or/and activity since metabolites have various functions including stimulatory and inhibitory effects on enzymes. The outcome of this change though was not depicted in the growth of the cells. The alterations in metabolism are a temperature related result but the cell seems to be able to compensate for the affected functions and ultimately maintain a steady state at 19 °C and 26 °C.

All strains showed a longer exponential phase at 13 °C compared to 19 °C and 26 °C which means that *L. aporus* needed more time to reach the stationary phase at lower temperatures. Having lower growth rate at 13 °C it is expected that it would take longer for the cells to deplete nutrients. Cell numbers followed an opposite trend compared to biovolume (high cell density, low biovolume and vice versa) leading to an assumption of a probable uniform biomass across the strains. But that was not the case. It has been suggested by Nanjappa (2012) that there is a trade off in biomass build up and time required for it in *L. aporus* since he observed higher biomass at low temperature. But the current results, combined with several of his, do not support this idea In fact some strains showed almost equal biomass in all three temperatures or even a much higher biomass at 19 °C; considering that the later were autumn strains, one could suggest that they were adapted for a higher biomass at 19 °C. On the other hand, the winter strains seem to be adapted for a higher biomass at the lowest temperature while no certain conclusion can be derived for the summer strain.

The growth experiment mainly aimed at detecting any possible differences in the response to different temperatures of strains isolated in different seasons. A pattern suggesting adaptation could be observed in biomass values with the most prevailing one being the adaptation of autumn strains at 19°C. Other than that, strains behaved in a more independent way suggesting genotypes with different physiological plasticity.

However, another interesting point from the results of the *L. aporus* growth experiments was a possible in-culture evolution in some of the strains used. Although during the present experiments the acclimatization time was the same for all strains, it should be kept in mind that three out of the six strains were taken from the SZN collection (B651, B704, B764). These strains were preserved at the same temperature (19 °C) for at least three years, a considerable long time. In the paper of Lakeman et al. (2009), a question is raised whether the observed differences in studies of intra-specific algal diversity are due to (a) adaptation to the natural environment from which strains were isolated or (b) evolutionary changes while being in culture. As it is stated in the same paper “the processes of evolution know no bounds, and do not cease to exert their influence even in our controlled laboratory environments”. This means that older cultures might depart more from their “natural state” than newer isolates. In my experiments, two of the three old SZN collection strains needed more time to reach a more stable growth rate at 26 °C during acclimatization, while six out of the nine growth rates that did not reach significant stability during acclimatization (Table 2.3.1) belonged to old strains. Finally, at 26 °C, all three old strains showed discrepancies between the experimental growth rates and the acclimatization ones. The odd behavior of the old strains compared to those recently isolated could provide an indication of in-culture evolution during the three years of culture maintenance, eventually leading to divergence from their original state.

Maintaining strains of the asexual species *L. aporus* in stable laboratory conditions should not be different from experimental evolution studies where mutations are the main evolutionary driver. Organisms such as bacteria and microalgae allow easier and faster investigations on evolutionary processes, because of their rapid generation times and small physical size. In microbial evolution

experiments, populations are placed in a new environment and allowed to adapt while this environment remains stable. Heterozygosity can increase due to asexual reproduction (due to the reduction of allele segregation) offering some fitness advantages over sexual reproduction, which can produce homozygotes of deleterious alleles. In such experiments with asexually reproducing populations, selection acts on standing variation between clonal lineages and heritable change within clonal lineages attributed to novel genetic and epigenetic mutations in order to restore any possible fitness loss (Lohbeck et al., 2012; Jin et al., 2013). Therefore, fitness measures such as growth rate are then used to compare the evolved genotypes with either their own ancestor or an evolving control population kept in a control environment. In a long-term microbial evolution experiment, Barrick et al. (2009) documented how fitness of replicate *E. coli* population initially rose and then leveled off even though genomic evolution continued, with adaptive mutations arising regularly in a stable environment.

In addition, organismal responses of strains growing in laboratory conditions for many generations lack effects from ecological interactions, such as competition for nutrients, grazing and viral attack. Collins (2011) found that excluding competitors for about 300 generations of growth in the alga *Chlamydomonas reinhardtii* limits its adaptive response to abiotic change. In the case of *L. aporus,* from 2010 to the time of the experiment, the old strains would have been growing in the same environment, under stable temperature, light conditions and absence of any kind of competition or threat, for more than 1,000 generations. Selection and adaptation under culture conditions could have changed the physiology of algal populations, mainly in the direction of a plasticity reduction, whereby genetic variations brought about by mutations, leading to possible genetic variability within an originally clonal strain, would not be favoured (not positively selected) in a stable environment, while epigenetic changes could have been lost over multiple generations. By contrast, under normal conditions (e.g. all strains been recently isolated) the differences in cell physiology would reflect plastic physiological responses, including epigenetic changes or other gene regulatory responses. In the same Lakeman's et al. (2009) paper, some recommendations were made regarding problems deriving from in-culture evolution, including

the minimization of the time between the isolations or performance of the experiments when culture are the same "age" in terms of generation, the application of identical culturing conditions and the use of sub-clones or replicate lines in the design of comparative growth experiment. In agreement with these considerations, the results for the more recently isolated strains 1A1, 1089-10 and 3A6 (2014 strains) of *L. aporus* were comparable, as well as B651 and B764 (2010 strains), whereas B704 (2009 strain) stands alone. The duplicates used for *L. aporus* during acclimatization behaved in the same exact way, but this only ensures that no evolutionary processes took place during the acclimatization period of the current experiment.

Summing up, three basic points emerge from the experiments conducted in this study:

- In *L. danicus* and *L. aporus,* temperature influences the cellular functions in an important degree, causing many physiological changes. Although the *L. danicus* growth experiment was not completed, the failure of acclimatization mainly at the high temperature condition in this species confirmed the lower tolerance, compared to *L. aporus*, to elevated temperatures (Nanjappa, 2012) which partly explains the contrasting seasonality of the two species. Of course, as it is already well understood by the scientific community, all environmental factors contribute to the specific responses of a species in nature and we should never neglect or overestimate the effect of an isolated stress factor studied in the laboratory.
- The long term maintenance of some of the *L. aporus* strains in culture might have had a strong effect on their growth response due to in-culture evolution, affecting their final behavior at the different temperatures. Because of that, it is hard to assign the reaction of the old strains entirely to their individual natural attributes since, when compared to the more recently isolated strains, in-culture evolution might have actually diverged them at a higher degree than it would be based on the differences of their natural state alone.
- In a first attempt to answer the question about the strains' responses to different temperatures and whether they respond differently due to their physiological plasticity or rather they represent distinct populations adapted to different conditions., *L. aporus*

strains seem to be mainly strains with a high diversity in physiological plasticity, and possibly high genetic diversity, and the same may stand for *L. danicus*. Thus, the diversity in the responses at the growth experiment of *L. aporus* can be described as genotype-specific trait responses, which would be a result of an extended range width of the encountered environmental conditions (Reusch and Boyd, 2013). Significant differences in growth rates among strains of a species have been noted in many studies (Rynearson and Armbrust, 2004; Whittaker et al., 2012). In a community-wide study by Boyd et al. (2013) the intraspecific variation in the diatoms *Thalassiosira rotula, Thalassiosira pseudonana* and the dinoflagellate *Akashiwo sanguinea* was higher at low and high temperature extremes suggesting that genotypic selection pressures have the largest influence under these conditions. Substantial strain (genetic) variability within a species can facilitate the success and widespread distribution of species such as *L. aporus* and *L. danicus*. Indeed, it has been repeatedly observed in physiology experiments such as in studies on effects of acclimatization time that adaptive phenotypic plasticity is widespread among phytoplankton species, acting as a form of phenotypic buffering (Bradshaw, 1965; Schlichting and Pigliucci, 1996; Pigliucci 2005). Such a capacity ensures the maintenance of functions across a broad range of environmental variables like temperature. At the same time, the presence of an autumn adapted population cannot be excluded assuming that both scenarios of physiological plasticity and adaptation can co-occur. Indeed, Jen-Pan Huang (2015) mentions that the immediate response of organisms to environmental change can involve both acclimatization based on phenotypic plasticity and adaptation based on selection.

2.4.1. Conclusion

Although intraspecific plasticity appeared to be the main driving force of *L. aporus* and *L. danicus* response to the different temperatures, the presence of adapted population cannot be excluded since clues pointing to this direction were also obvious. Nevertheless, the overall results of the experiments conducted on *Leptocylindrus* highlight the importance of using many strains for a

given region or season, as one, two or even three strains might not be reliable representatives of a species or of a local/ seasonal population. This conclusion has to be considered especially when interpreting gene expression of individual species and transcriptomic responses at different temperatures. The topic of intraspecific variability will be further discussed in other chapters of the thesis.

Chapter 3. *L. aporus* gene expression changes in response to different temperatures

3.1. Introduction

As already mentioned in Chapter 2, temperature is one of the most important factors regarding the growth response of species with diatoms' exact relationship to it being still unclear. In any case, it is generally accepted that temperature plays an essential role in the determination of the spatial and temporal distribution of species. Furthermore, the adaptive potential of diatom populations (and phytoplankton in general) is little understood. Likewise, the role of adaptation as one of the principal mechanisms driving speciation and the response of species and communities to environmental change remains poorly known. Investigating the actual reaction regarding the gene expression of a diatom species to different temperatures could shed more light into this dark area. Up to date, studies on diatoms' adaptation or acclimation potential to different temperatures have taken advantage mainly of real time PCR techniques as well as EST-macroarrays, (Mock and Hock, 2005; Parker and Armbrust, 2005; Bayer-Giraldi, 2010; Helbling et al., 2011; Kinoshita et al., 2001) whereas RNA sequencing has not yet largely been utilized (Koester et al., 2013; Toseland et al., 2013).

Plants and other organisms have an inherent ability, called basal thermotolerance, to survive at temperatures above the optimal for growth and an ability, called acquired thermotolerance, to acquire tolerance to otherwise lethal heat stress. The latter one is induced by a short acclimation period at moderately high yet bearable temperatures (Kapoor et al., 1990; Flahaut et al., 1996; Larkindale et al., 2007). Conversely, stress tolerance can be induced by exposure to reduced temperature; the ability to tolerate low temperatures without damage is known as chilling tolerance whereas the enhanced tolerance to the physical and physiochemical changes of freezing stress is cold acclimation (Somerville, 1995; Thomashow, 1999). Low or high temperatures close to the environmental extremes experienced by the organisms might act as stress factors leading to activation of stress related pathways engaged in restoring cellular homeostasis. The stress response is a universal and highly conserved mechanism of cell survival to elevated or decreased temperature and other unfavorable environmental conditions (Lindquist, 1986) and it is characterized by an increase in stress inducible proteins coupled with a decrease in constitutive

protein production (Schlesinger et al., 1982). In thermo-intolerant diatoms there is a lag period between heat shock and stress protein synthesis (Rousch et al., 2004). Temperature stress changes the structure, catalytic properties and function of enzymes and membrane metabolite transporters so at first the organism will try to adjust its cellular metabolism through regulatory mechanisms and restore normal metabolite levels and metabolic fluxes (Schwender et al., 2004). Secondly, enhanced tolerance mechanisms involving extensive reprogramming of gene expression and further modifications of metabolism are triggered. Heat and/ or cold stress have a complex impact on cell function proving that many processes are involved in temperature tolerance:

- Membrane-linked processes are affected due to alterations in membrane fluidity and permeability through changes in lipid composition (Sangwan et al., 2002; Welti et al., 2002). Adaptation of the lipid membranes is a common low-temperature adaptation mechanism in the sea ice diatoms isolated from the Antarctic (Mock and Kroon, 2002) and Arctic ice shelves (Henderson et al., 1998).
- Imbalanced metabolic pathways or complete enzyme inactivation due to alterations in enzyme activity and protein denaturation (Kampinga et al., 1995). The level of ubiquitinated proteins and therefore the protein degradation rate has been found increased during heat shock in *Skeletonema costatum* (Scoccianti et al., 1995). Reduction of photosynthesis lies within these consequences.
- Heat/ cold - induced oxidative stress caused by the production of reactive (also called active) oxygen species due to membrane and protein damage (Larkindale and Knight, 2002). Reactive oxygen species accumulate in plant cells during various abiotic stresses and it has been shown to have a strong influence on regulation of gene expression as well (Lee et al., 2002).
- Activation of specific signaling pathways linked to stress-response. The signal resulting from the temperature stress is transduced downstream and many signaling pathways are activated; the components of these pathways are various and can be calcium, reactive oxygen species, protein kinases (mitogen-activated protein kinase, MAPK, involved in

osmotic stress signaling as well), protein phosphatase and lipid signaling cascades. The final target could be stress-responsive genes or transcription factors that regulate expression and function of genes (heat/ cold regulated genes), which ultimately leads to adaptation and survival during unfavorable conditions (Yamaguchi-Shinozaki and Shinozaki, 2006).

In photosynthetic microorganisms, including diatoms, the same molecular responses are followed; adaptation and acclimation to temperature shifts include the maintenance of membrane fluidity (Mock and Kroon, 2002; Ralph et al., 2005), photosynthetic electron transport and energy balance (Lomas and Gilbert, 1999; Parker and Armbrust, 2005), and the activation of heat/ cold-adapted enzymes (Loppes et al., 1996; Salvucci and Brandner, 2004).

Heat Shock Proteins

The effector genes encoding proteins regulated by temperature include chaperones and in particular heat shock proteins (HSPs), one of the best characterized aspect of acquired thermotolerance.

Diverse physiological stresses like heat or cold produce multiple changes in a cell that ultimately affect protein structures and function. *HSP*s were initially characterized as a highly conserved family of genes whose expression was induced by heat shock but now it is well known that HSPs play a prominent role in many of the most basic processes of the cell beyond response to heat (Ritossa, 1996). *HSP*s are molecular chaperones that facilitate the synthesis and folding of proteins but also participate in protein assembly, turn-over, export and regulation. In addition, HSP chaperones are involved in the repair of stress-accumulated misfolded proteins preventing their aggregation (Hartl, 1996; Schmitt et al., 2007). In fact, diverse stresses, including heavy metals, oxidative stress and cold, induce the expression of *HSP* genes (Colinet et al., 2010; Miura and Furumoto, 2013). In plants, HSP expression can be induced by cold (Timperio et al., 2008). During development, growth and adaptation the synthesis of HSPs is regulated by Heat Shock Transcription Factors (HSFs) which are not strictly heat inducible (Wu, 1995; Kim et al., 1997).

HSFs play a key role in signal transduction pathways involved in the activation of genes in response to numerous types of environmental stress (Nishizawa-Yokoi et al., 2011).

In diatoms, the role of HSPs during stress response has also been confirmed; thermal stress considerably upregulates HSP90 in the Antarctic diatom *Chaetoceros neogracile* (Jung et al., 2007) and in *Ditylum brightwellii*, along with HSP70, (Guo et al., 2013) whereas a shift to colder temperature upregulates HSP70 in the polar diatom *Fragilariopsis cylindrus* (Mock and Valentin, 2004). In addition, HSPs were found to behave as signals of senescence/ aging and short-term exposure to stress in *Skeletonema marinoi* (Lauritano et al., 2015).

Transposable elements

In addition to HSPs, transposable elements have often been shown to be linked to various biotic and abiotic stressful conditions such as temperature changes (Bouvet et al., 2008; Maumus et al., 2009; Rakocevic et al., 2009).

Transposable elements (TEs) are short mobile DNA sequences, with diverse structures, able to move within the genome. They contain a high number of repetitions and short sequences. They are often gathered in non-coding regions of DNA such as the heterochromatin and for this reason they were known as “junk DNA” for a long time since their discovery by McClintock (McClintock, 1951; Xiong and Eickbush, 1990; Hermann et al., 2014). Their vast diversity and ubiquitous presence in living organisms led to many extensive studies on their involvement in genomics, thanks to which we now know a lot more about the large world of TEs but still not enough to fully understand their role in ecology and evolution of species. Due to their diversity and the absence of a universal structural rule for their identification, TEs are usually identified based on their similarity and/or homology with already established TEs, thus any TEs with novel structures or several mutations are difficult to identify and classify (Hermann et al., 2014). Further information on the structure and classification of transposons can be found in the Appendix.

Active TEs are highly mutagenic when targeting protein coding genes for insertion, also causing chromosome breakage, genome rearrangement, illegitimate recombination, altering splicing and polyadenylation patterns of neighboring genes and functioning as promoters or enhancers (Girard

and Freeling, 1999). In plants, the invasion has been so massive that TEs now make up most of their genome, obviously having affected the genome structure through chromosomal sequence rearrangements and disrupted gene expression (Chenais et al., 2012). However, deletions or mutations have inactivated most of the TEs (immobile remnants) whereas many TEs have lost the ability to move by themselves so they use the machinery of others to transpose and replicate (non-autonomous elements) (Feschotte and Mouches, 2000). In addition, genomes have evolved epigenetic 'defense' mechanisms to suppress the activity and the potentially harmful effects of TEs, producing in that way some full-length intact autonomous TEs silenced by a repressive chromatin environment (Slotkin and Martienssen, 2007). TEs can be reactivated by stress contributing in that way to the generation of the raw diversity that a species requires over evolutionary time to survive the specific stress. It is an adaptive response that functions as a long-term strategy to increase variability, not necessarily genetic, for selection (Capy et al., 2000). All active or reactivated TEs can produce genetic and phenotypic variability between individuals because of their polymorphic locations but also by subjecting nearby genes to the epigenetic regulation that is originally targeted for the TE (Rakyan et al., 2002). Based on these, one can conclude that TEs provide fuel for evolution, yet this is presumably not the aim of their transposition. In any case, the stress-induced reaction of TEs and their invasion in a new genome but also their control and epigenetic regulation might account for some key aspects of genome evolution (Kidwell and Lisch, 2000; Slotkin and Martienssen, 2007). TE-induced mutations have been associated with adaptation to the environment in numerous studies (de Visser et al., 2004; Stoebel and Dorman, 2010; Gaffè et al., 2011; Kanazawa et al., 2009; Chu et al., 2011; Gonzalez et al., 2010). An ever-changing environment may allow TEs to play an important role in the responsive capacity of their hosts by increasing the genome ability to cope with environmental challenges.

In diatoms, a specific example is *Phaeodactylum tricornutum.* Long Terminal Repeat (LTR, see Appendix 1 for classes of transposable elements) elements are so abundant in *P. tricornutum* that it is suggested that major genome rearrangements resulting from massive activation of LTR would

allow the organism to respond rapidly to changing environmental conditions (Maumus et al., 2009). As already mentioned, TE activation can be triggered by or in response to environmental stress. The question is how exactly TEs can be activated or reactivated by stress?

- Some TEs are sufficient alone to activate TE transcription in response to stress. For instance, the *Blackbeard* retrotransposon in *P. tricornutum* is activated under nitrate starvation stress through hypomethylation providing a link between environmental stress and chromatin remodeling in diatoms (Maumus et al., 2009).
- Other TEs are activated through their regulatory sequences. Several LTR retrotransposons contain *cis*-regulatory elements that have been found to be similar to the motifs required for the activation of stress-responsive genes and so they trigger transposon expression in response to the particular stimulus (Kumar and Bennetzen, 1999; Grandbastien et al., 2005). In *Arabidopsis thaliana* a LTR-copia type retrotransposon acquired a heat-responsive element recognized by plant derived heat stress defense factors (HSFA1 and HSFA2) resulting in a specific response under elevated temperatures (Cavrak et al., 2014).
- TEs might integrate close to stress-responsive genes and hence activated along with them. Certain TEs can show a strong preference to integrate close to promoters induced by specific stress conditions such as heat, implying a unique response that leads to specific expression level alteration of the stress responsive genes (Guo and Levin, 2010).
- TEs can be activated by a secondary response to the specific initial stress. Other cellular mechanisms affected by the stress trigger the activation of the TE (Zhu et al., 2003; Dai et al., 2007).

The effects of TEs as novel promoters or enhancers for nearby genes should not be neglected as they might as well provide the surrounding genes the capacity to respond to certain stress situations (Naito et al., 2009; Makarevitch et al., 2015).

In diatoms, the investigation of TEs could provide a model to evaluate their possible role in the great diversity and adaptation capacities of these organisms worldwide (Hermann et al., 2014). The level of TE invasion in diatoms is considerably low, and mostly related to the LTR

retrotransposons superfamily, which make up 90% of TEs in *P. tricornutum* and 58% in *T. pseudonana* (Maumus et al., 2009). Phylogenetic analysis of Ty1-Copia (LTR retrotransposon) conserved reverse transcriptase domains from the genomes of those two species revealed the existence of seven groups (six of them being diatom specific) called CoDi1 to CoDi7. The groups were divided into three lineages; one consists of elements mainly found in *P. tricornutum,* the second is made up of elements found in both species and third, Copia lineage CoDi6, includes elements again from both species but also closely related to Ty1-Copia from other organisms, mostly marine. Two of the diatom specific retrotransposons, namely *Blackbeard* and *Surcouf,* were found active and overexpressed during nitrate starvation (Maumus et al., 2009). *Surcouf* was also overexpressed at high temperature just as small Heat Shock Proteins (sHSP). In fact similar sequences were detected in both the sHSP promoters and the 5' DNA sequence of *Surcouf* but not of *Blackbeard*, suggesting that sHSP and *Surcouf* may be partly co-regulated (Egue et al., 2015). *Surcouf* and possibly *Blackbeard* might therefore act as environmental sensors in the response of the diatoms to stress. Moreover, the expression of *Blackbeard* is suggested to be under epigenetic control, which implies a direct link between environmental stress and chromatin modeling in the processes of adaptation to environmental variations (Maumus et al., 2009).

In the current study, three conditions (at 13 °C, 19 °C and 26 °C) have been chosen in order to investigate the functional diversity of *L. aporus* in response to temperature. A first idea on the species behavior has been already acquired in Chapter 2 where it was suggested that the species growth is equally efficient at medium and high temperature but less at low temperature conditions. In the same chapter, the first evidences on intraspecific diversity and/or adaptation to culturing conditions were noted. Here, those results were also kept in mind while the expression patterns of *L. aporus* at the three temperatures were compared and the genes found significantly differentially expressed were further investigated for specific functions related to the ones mentioned above. In particular, greater attention was given to genes encoding for HSPs and TEs. The RNA-seq results for the genes of interest have been validated on the same biological samples under the same conditions by a second independent method, while as a further exploration of

their role in stress response, the expression of the selected TEs and HSPs was investigated also in samples with a shorter acclimatization time period at the low and high temperature. Quantitative real-time polymerase chain reaction (qRT-PCR) is one of the techniques mostly used for fast, sensitive and accurate gene expression analysis and the one usually chosen for validation of RNA-seq results (Siaut et al., 2007). In the end, three HSP and five TE related transcripts were selected for validation and exploration analysis with the qRT-PCR technique. So far, information on HSPs in diatoms is restricted and their role not fully understood while results on TEs are limited only to two diatom species. For this reason, it is important to investigate more diatom species and understand whether HSPs are involved in heat/ cold stress responses and if TEs do indeed play an essential role in diatom responses to environmental changes and ultimately if they are responsible for the genetic diversity that have allowed diatoms to adapt so successfully to so many environments (Bowler et al., 2008; Maumus et al., 2009; Hermann et al., 2014).

3.2. Materials and Methods

3.2.1. RNA-sequencing and downstream analysis

1. *RNA extraction and sequencing*

Based on the results of the molecular characterization, strains of *L. aporus* were selected so that, in combination with strains already available in the SZN collection, they would cover as high diversity as possible in terms of seasonality and physiological characteristics (Table 3.2.1.1). Intra-species molecular similarity was also checked and it was found that all strains selected were identical regarding ITS marker.

Table 3.2.1.1 *L. aporus* strains selected from the isolated and SZN collection strains for RNA extraction and sequencing.

#	Isolation Date	Strain Code	Species	Temperature
1	21/8/2010	B651	*L. aporus*	13 °C
				19 °C
				26 °C
2	20/12/2013	1A1	*L. aporus*	13 °C
				19 °C
				26 °C
3	28/01/2014	3A6	*L. aporus*	13 °C
				19 °C
				26 °C

L. aporus strains were acclimatized to three temperatures, 13 °C, 19 °C and 26 °C, at a light intensity of 100 µmol photons m^{-2} sec^{-1} and a photoperiod of L:D, 12:12. Before reaching the final temperatures, cultures spent one week at intermediate temperatures, 15 °C and 23 °C (see Fig. 2.2.2.1 in Chapter 2). Cultures were kept in 100 ml of K + Si medium and waterbaths were used for regulation of temperature. Cells were counted and chlorophyll fluorescence was measured daily using a Turner 10-005 fluorometer. When cultures reached a concentration of 2,000 cells/ml while at their exponential phase they were transferred to 1 liter (1L). Growth rate was calculated based on plots of the logarithmic values (base 10) of fluorescence and days in the same way described in Chapter 2. RNA was extracted as follows:

1. Cultures were grown until a final concentration of 50,000 cells/ ml in 1 L while it was also ensured they were at exponential phase. Then they were harvested by filtration on 47 mm MF-Millipore mixed cellulose membrane filter (1.2 µm pore size).
2. The filter was cut into two halves, each half stored in 2 ml Eppendorf tube.
3. 1.5 ml of TRIzol Reagent was added to each Eppendorf and vortexed briefly. One tube was flash frozen in liquid nitrogen and then stored at -80^{o}C while the other one was immediately used in the next extraction step.
4. Glass beads were added and then the tube was put on thermoshaker for 10 min at 60^{o}C and maximum speed.
5. Brief centrifugation of one minute followed. Supernatant was transferred into a new 2 ml Eppendorf tube.
6. 300 µl of chloroform was added, vigorous shake and incubation at room temperature for 15 minutes followed.
7. Centrifugation of 15 minutes at 4 °C and 10,600 rpm (12,000 x g) followed. The upper most aqueous phase containing RNA was transferred into a new 1.5 ml tube. Step 6-7 were repeated adding as much chloroform as the supernatant obtained.
8. Equal volume (approximately 750 µl) of isopropanol was added and the tube was inverted to mix. Then it was incubated at 4 °C for one hour to overnight.

9. Centrifugation of 10 minutes at 4°C and 10,600 rpm (12,000 x g) and removal of supernatant followed.
10. 1.5 ml of 75% ethanol was added, tube was gently inverted to wash the pellet and stored at -20 °C overnight or more.
11. Centrifugation of 10 minutes at 4°C and 10,600 rpm (12,000 x g) and removal of supernatant followed. The pellet was dried for 15-20 minutes.
12. The pellet was resuspended in 44 µl of DEPC water and incubation at 55-60°C for 5-10 minutes.

DNase treatment

Roche DNase I recombinant, RNase-free, (10 u/µl) and 10x incubation buffer was used. 1 µl of enzyme solution and 5 µl of buffer were added to 44 µl of RNA sample and incubated for 10 minutes.

RNA cleanup

In order to clean up the RNA and remove DNase treatment reagents, the QIAGEN RNeasy Mini Kit was used. After clean up, RNA was run on Agilent Bioanalyzer 2100 RNA Nano LabChip (Agilent, Palo Alto, CA) for qualitative control. Good quality RNA samples were sent for sequencing. Within the framework of the collaboration of SZN with EMBL Genomics Core Facilities, an offer for preparation of RNA sequencing libraries and deep sequencing on the Illumina HiSeq instrument (pair-end, 50 bases) was provided. All *L. aporus* samples were placed in the same lane.

2. <u>Bioinformatics downstream analysis</u>

The bioinformatics analyses to assemble and annotate the transcriptome and to identify the differentially expressed transcripts were performed by Remo Sanges and Francesco Musacchia (SZN). The reads produced from the RNA sequencing were processed as follows:

1. Quality check and trimming. The quality of the reads was checked using FastQC v0.11.3 (Babraham Bioinformatics) with the default parameters and the trimming of adapters was performed by Trimmomatic 0.33 (Bolger et al., 2014) where the sliding window was kept low in order to permit a reliable quality average. Reads that were half the initial sequence

length (50 nucleotides) were kept. After trimming, one more FastQC check was run in order to confirm the quality improvement.

2. Assembling. Paired-end reads were assembled with Trinity 2.0.6 (Grabherr et al., 2011). CD-HIT-EST (Li et al, 2001) was used in order to reduce the redundancy of isoforms that belong to the same gene.
3. Quantification and filtering. The reads were aligned to the transcripts generated by the assembler with Bowtie v.1.1. (Lakeman et al., 2012) and SAMtools (Li et al., 2009) were used to count the mapped reads. EdgeR (Robinson et al, 2010) and faSomeRecord function from UCSC Genome Browser (Kent et al., 2002) were used in order to calculate CPMs (Count Per Million) and extract transcripts with a CPM greater than 1 for at least 2 samples, respectively.
4. Annotation. Annocript 1.1.2 (Musacchia et al., 2015) was executed for the annotation of the transcripts using Swiss-Prot (Bairoch and Apweiler, 1996), UniRef90 (Suzek et al., 2007), the Conserved Domains Database (CDD, Marchler-Bauer et al., 2014), Rfam (Griffiths-Jones et al., 2003) and SILVA (Quast et al., 2013) database.
5. Differential expression analysis. Annocript uses EdgeR to evaluate the expression of the transcripts obtained and statistically distinguishes the expression of the genes among the experiments. The following comparisons were performed: 13 °C-19 °C, 19 °C-26 °C, 13 °C-26 °C. Transcripts were considered differentially expressed when FDR (False Discovery Rate) ≤ 0.05 and FC (Fold Change) > 2. A k-means cluster analysis based on unsupervised classification was applied to the differentially expressed genes. The clustering was performed in MultiExperiment Viewer (MeV) and the parameters were: k=15 (number of clusters), 500 clusters re-ordering iteration, Pearson uncentered as distance measure and calculation of medians (Howe et al., 2011).
6. Enrichment analysis. The Gene Ontology (GO) terms that were enriched in the list of the differentially expressed transcripts in respect to the overall transcriptome were identified by the R prop.test function (R Core Team, 2015). The parameters used were 10 as the

minimum number of transcripts associated to a GO class and a cut-off of adjusted p-value equal to 0.1 in order to consider a class significant.

Part of the visualization of the results was done with T-Rex (RNA Expression Analysis) webserver for RNA-seq expression data (de Jong et al., 2015).

3.2.2. qRT-PCR analysis

1. *Selection of target and reference genes*

After analyzing the RNA-seq results, specific transcripts of interest were selected for validation with quantitative real-time PCR (qRT-PCR). To achieve this aim, TE-related transcripts were mainly selected, along with some other temperature-related transcripts. In order to select the best candidates, the nucleotide and protein sequences of the significantly differentially expressed transposon-related transcripts were derived and 'blated' (tblastx and blastp respectively) against the diatom genomes available in Ensembl (*Phaeodactylum tricornutum, Thalassiosira oceanica, Thalassiosira pseudonana).* BLAT is an alignment tool like BLAST, but it is structured differently. BLAT is commonly used to look up the location of a sequence in the genome since the target database of BLAT is not a set of GenBank sequences, but instead an index derived from the assembly of the entire genome. BLAT of DNA is designed to quickly find sequences of 95% and greater similarity of length 40 bases or more. Based on the results, transposon-related transcripts were grouped and then used to build a phylogenetic tree. The nucleotide sequences were also imported in RepeatMasker, along with the CoDi sequences from Maumus et al. (2009). RepeatMasker is a program that screens DNA sequences for interspersed repeats and low complexity DNA sequences (Smit et al., 2015).

What is very important in qRT-PCR is the selection of appropriate reference genes in order to normalize the expression ratios. So in addition to the target genes to be validated, reference genes were also selected for the qRT-PCR analysis. Reference genes are genes whose expression should remain stable through the different conditions tested in the experiment and they act as a control for the calculation of the target gene expression. In this case histone H4, tubulin beta-6 chain (TUBB), tubulin alpha 1A (TUBA) and TATA box were selected and checked for their stability

using NormFinder software, an algorithm for identifying the optimal normalization gene among a set of candidates (Andersen et al., 2004). These genes are commonly used as reference genes in qRT-PCR and were found stable in other diatoms (Nanjappa, in prep; Adelfi et al., 2014). The region of the transcripts to be amplified was selected based on the following criteria:

1. The region should be included in the open reading frame (ORF) of the transcript. In order to ensure that the nucleotide sequence of the transcript was translated to a protein sequence using the ExPASy translate tool which is available online.
2. The region should be a conserved domain if possible. This was checked based on the annotation by Annocript.
3. The length of the amplicon should not exceed 200 bp (usually 60-150 bp).

Based on these criteria primers were designed on the Primer-Blast tool of NCBI. The pair of primers that were more suitable in terms of GC content (50-60%), length (min: 18 °C, max: 24 °C), melting temperature (min: 60 °C, max: 63 °C, best: 60 °C) and 3'-complementarity and self-complementarity scores (ranges from 0-3 and 2-6 respectively) were preferred (Table 3.2.2.1).

Table 3.2.2.1 Reference and target genes and their corresponding primers. The related selection criteria are presented for each pair.

Gene/Domain	Transcript ID	Sequence (5'-3')	Primer Length	Ampl. Length	Tm	GC %	Self-compl.	Self-3'-compl.
Histone 4	H4	Forw.: TCGTGGTGTTCTCAAGGTAT Rev.: TTTCCCTGCCTCTTCAAGGC	20 20	122	56.19 60.25	45 55	2 5	5 4
Tubulin beta	TUBB	Forw.: GGTAGAGAACGCGGACCAAT Rev.: TTGTCCGGGGAACCTCAAAG	20 20	160	59.82 59.59	55 55	4 4	2 1
Tubulin alpha	TUBA	Forw.: TCAATTCGGGACAGTGCCTC Rev.: GCCAATGTTCCTGGTGGAGA	20 20	95	60.04 59.96	55 55	5 3	3 0
TATA box	TATA	Forw.: TGACAGTGCCAACGGGTATC Rev.: CCCGTAGCCTTTGGTTCCAT	20 20	217	59.82 59.82	55 55	4 4	2 1
RNase H RT Ty1	TR6356	Forw.: GTAGCACGGAGGCGGAATTA Rev.: GCTCTTTCTCCCGTTCGTCT	20 20	158	59.9 59.76	55 55	4 2	2 0
RNase H RT Ty1	TR7186	Forw.: CGCGAGTCATGCCACTAATC Rev.: AACCCCAGTCCTTAATGCCA	20 20	154	59.14 58.92	55 50	4 4	3 3
Reverse Transcriptase	TR6586	Forw.: CACTCGATGCAAGCAAGTCG Rev.: CCCCTTTGATGAGTGCGTCT	20 20	80	59.91 60.04	55 55	4 2	2 0
Integrase core (Ty3 Gypsy)	TR6506_i3	Forw.: TGGCCGAAGTACAGGACCTA Rev.: ATTGGCCTGAGGGTTTCGAG	20	80	59.96 60.04	55 55	5 5	3 2
Integrase core (Ty3 Gypsy)	TR6506_i5	Forw.: AGAGAGCGGACGAAATAGCG Rev.: ACAATTACGTGCTGAGGCCA	20 20	76	59.97 59.96	55 50	2 4	2 2
Heat Stress Transcription Factor A-1a	HSFA	Forw.: ACCATGGGGCAACCAAGATA Rev.: GTGGGGAGATTTCGGCCATT	20 20	121	58.99 60.4	50 55	6 4	2 1
Heat Stress Transcription Factor B-2a	HSFB	Forw.: GTCGTCGTTTCGTAAGCAGC Rev.: CAGCTTGGGCATTCCTCGTA	20 20	109	59.91 60.11	55 55	4 4	3 2
Stress-inducible yeast Mpv17 (SYM1)	SYM1	Forw: TGTTGGGGTATATGGATACCAGT Rev.: TTCGGAGAAACTCTGGAACAA	23 21	78	58.43 57.17	43.48 42.86	6 3	2 0

2. qRT-PCR analysis

For the qRT-PCR validation the following set of *L. aporus* RNA samples was used:

B651, 1A1, 3A6 (acclimatized in 2015). The same three strains that were used for RNA-seq were also used for qRT-PCR validation. However, the samples were not the same as the one sequenced but they were duplicates that had been acclimatized in the same period and for approximately the same time as the ones sent for sequencing (100 – 150 days). In the following, the RNA-seq samples will be named with an additional **" _r "** in the end of their IDs while samples from the same period used for qRT-PCR will be named with the **"2015"** addition in their names.

In addition, in order to further explore the behavior of the target genes in response to different acclimatization and cultivation times, qRT-PCR experiments were also performed on two different sets of RNA samples:

i. **B651, 1A1, 3A6 (acclimatized in 2016).** The same strains mentioned above were acclimatized again in 2016 for half the original time (40 – 60 days) and used for qRT-PCR validation. These will be referred to as the "**2016"** or **exploration set 1** samples (Table 3.2.2.3).

ii. **1188A1, 1189A3, 1189B3**. Three strains that were recently isolated and characterized based on ITS in the same way mentioned in section 2.2 were acclimatized for an even shorter time (10 – 40 days) and then used for qRT-PCR. These will be referred to as the **exploration set 2** samples (Table 3.2.2.2 and 3.2.2.3).

Table 3.2.2.2 Strains isolated in 2016 and used in the qRT-PCR validation of the selected transcripts.

New *L. aporus* strains	Isolation Date
1188A1	04/02/2016
1189A3	19/02/2016
1189B3	19/02/2016

Table 3.2.2.3 Isolation, start and end of acclimatization dates and acclimatization duration for each sample used in qRT-PCR.

Validation Set	B651_2015			1A1_2015			3A6_2015		
Temp.	**13**	**19**	**26**	**13**	**19**	**26**	**13**	**19**	**26**
Isolation Date	21/08/2010			20/12/2013			28/01/14		
Start of Acclim.	01/09/2014			01/09/2014			01/09/2014		
End	21/01/ 2015	01/02/ 2015	24/03/ 2015	26/03/ 2015	16/12/ 2014	17/12/ 2014	23/01/ 2015	17/12/ 2014	16/01/ 2015
Duration (days)	140	150	203	205	105	105	142	105	135
Exploration Set 1	**B651_2016**			**1A1_2016**			**3A6-2016**		
Start of Second Acclim.	15/01/2016			15/01/2016			15/01/2016		
End	14/03/ 2016	25/02/ 2016	02/03/ 2016	21/03/ 2016	24/02/ 2016	29/02/ 2016	20/03/ 2016	01/03/ 2016	05/03/ 2016
Duration (days)	60	40	47	66	39	44	65	46	50
Exploration Set 2	**1188A1**			**1189A3**			**1189B3**		
Isolation Date	04/02/2016			19/02/2016			19/02/2016		
Start of Acclim.	04/03/2016			11/04/2016			15/04/2016		
End	17/ 04	21/ 03	08/ 04	07/05	19/04	28/04	07/05	24/04	03/06
Duration (days)	43	17	34	26	8	17	22	9	19

Here it should also be mentioned that 1189B3 was not able to grow to the same concentration as the rest of the strains during the acclimatization at 26 °C. Therefore a whole filter of 10,000 cells/ml was used which corresponds to less than half of the concentration of the other samples. In the end, as most significant differences in the differential expression analysis among temperatures were seen between the highest and the lowest temperature, the validation and exploration were performed between these two temperatures

The RNA was extracted in the same exact way as it was done for the RNA sequencing. RNA was analysed by gel electrophoresis (1% agarose w/v) and NanoDrop spectrophotometer to determine the quality as 260/280 nm and 260/230 nm absorbance ratios. Good quality RNA was reverse transcribed into cDNA using the QuantiTect Reverse Transcription Kit (Qiagen, Venlo, Limburgo, Netherlands). The primers' specificity was checked by blasting against the whole *L. aporus* transcriptome and by conducting a gradient PCR for a temperature range of 59.6 – 66.9 °C. The reactions were done in final volume of 10 µl: cDNA 1µl, forward primer (10 µM), reverse primer (10 µM), PCR reaction buffer with $MgCl_2$ (10x), dNTP (10x), Taq DNA Polymerase (5u/µl). The products were run on 1.5% agarose gel in order to specify the size of the amplicon and confirm a single band, which means a single region is amplified for each primer. The primers efficiency was calculated with an at least five point serial 10-fold dilution using the Standard Curve method of

ViiA™ 7 Real-Time PCR System (Applied Biosystems by Life Technologies, Carlsbad, CA, USA). The efficiency of all primers were ranging from 1.85 to 2.41. In addition, the melting curves, which ideally should show a single peak corresponding to a single product, were checked in order to verify the specificity of the primers seen already by the agarose gel.

Finally qRT-PCR was performed for all samples in triplicates with a negative control. qRT-PCR plates were set up according to the "standard SBM protocol":

Add Order	***Sample***	
1°	Fast SYBR Green Master Mix with ROX*	**5,0 ul**
2°	Oligo F + Oligo R [0,7 pmol/ul] each	**4,0 ul**
3°	cDNA	1,0 ul
	V tot.	10,0 ul

*Applied Biosystems by Life Technologies, Carlsbad, CA, USA

Reaction plates were analysed according to the following thermal profile:

Fast SYBR (duration ≈ 50 min)		
Step 1	95°-10 min	
Step 2	95°-1 sec	
Step 4	60 °-20 sec	
Go to	2	39 times
Step 6	Melting	

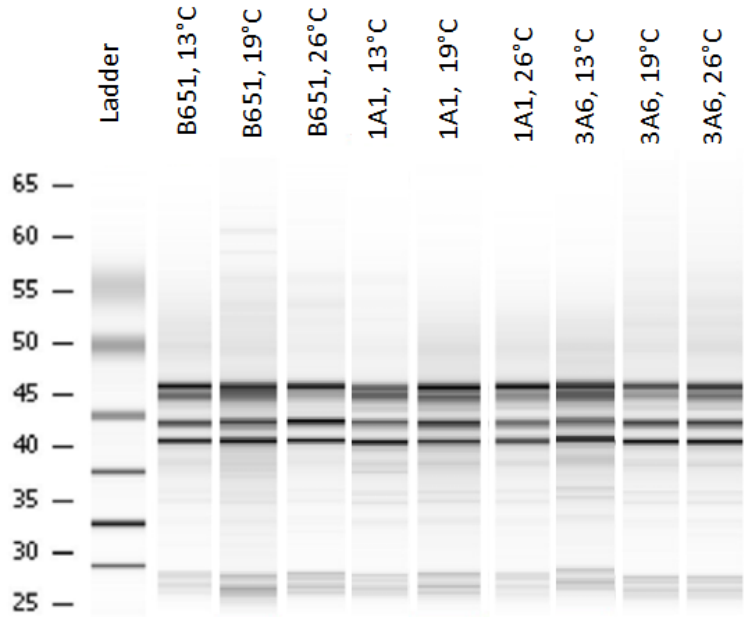
The analysis of the expression levels was performed using TUBA and TUBB as reference genes, since TUBB was detected as the best gene by NormFinder (smallest stability value: SV^{TUBA}=0.388, SV^{TUBB}=0.247, SV^{TATA}=0.301) and TUBA showed the lowest standard error (SE^{TUBA}=0.199, SE^{TUBB}=0.221, SE^{TATA}=0.202), and the Relative Expression Software Tool-Multiple Condition Solver (REST-MCS) which is a software for the calculation of the relative expression in qRT-PCR. The software uses the Pair Wise Fixed Reallocation Randomization Test (Pfaffl et al., 2002). The relative expression ratio is calculated from the real-time PCR efficiencies (E) and the crossing point (CP) difference (Δ) of an unknown sample versus a control ($\Delta CP_{control-sample}$):

$$\text{Ratio (R)} = (E_{target})^{\Delta CPtarget\ (control-sample)} / (E_{ref})^{\Delta CP\ ref\ (control-sample)}$$

The relative expression ratio (R) of the targeted genes was computed as the expression variation between high temperature samples, set as control, against the low temperature samples, set as condition, normalized over the expression variation of reference genes whose expression levels were not regulated in specific experimental conditions.

3.3. Results

A good quality RNA should have a high RNA integrity number (RIN ≥ 7) that is calculated based on the size of the two bands in the electrophoresis or the height of the two peaks in the electropherogram (bands/ peaks correspond to the 18S and 28S ribosomal subunits). In the case of diatoms one extra peak is present at a position close to 18S. Due to this unique nature of the diatom RNA, the 18S and 28S peaks are sometimes wrongly assigned by the Bioanalyzer and the RIN values can be incorrect. Therefore, for the assessment of the RNA quality only the electrophoresis results and electropherograms were used and according to them all samples were found of an acceptable quality for sequencing (Fig. 3.3.1).



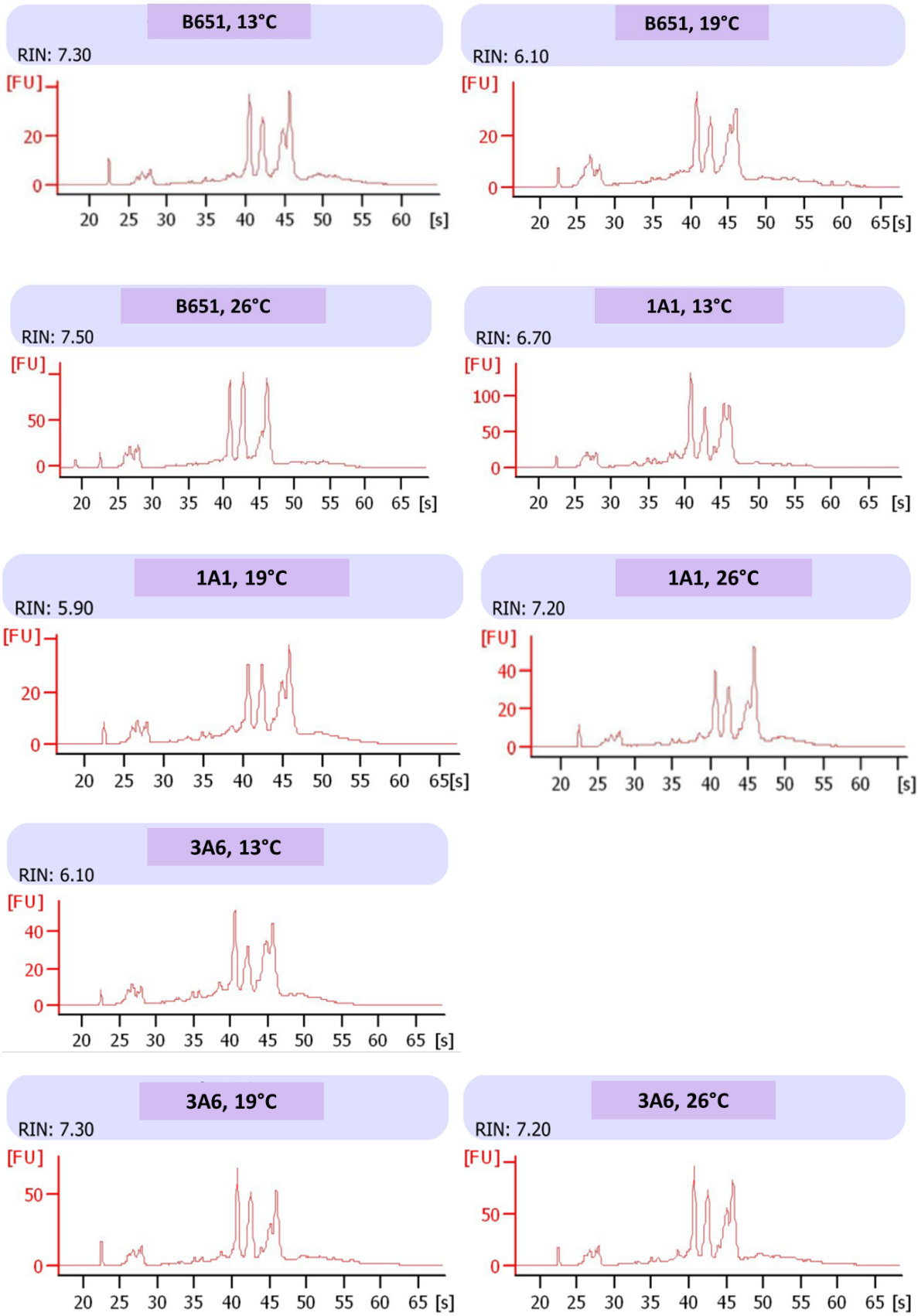


Figure 3.3.10 Bioanalyzer results of *L. aporus* RNA samples sent for sequencing. The electrophoresis (above) and electropherogram (below) results show the typical pattern of a diatom good quality RNA with the three expected bands/ peaks.

The final, complete transcriptome of *L. aporus* consisted of 19,963 transcripts. Further statistics on the transcriptome follow (Table 3.3.1).

Table 3.3.1. *L. aporus* transcriptome statistics.

***L. aporus* final transcriptome**	
Number of transcripts	19,963
Mean sequence length	1,315 bp
Minimum sequence length	224 bp
Maximum sequence length	11,566 bp
Swiss-Prot hits	7,129
UniRef hits	12,489
Domains hits	9,339
Ribosomal RNAs hits	127
Transcripts with at least one blast result	12,872 (64.48%)

The expression values (CPMs) of each sample showed an equal distribution for all strains, especially between same temperatures (Fig. 3.3.2).

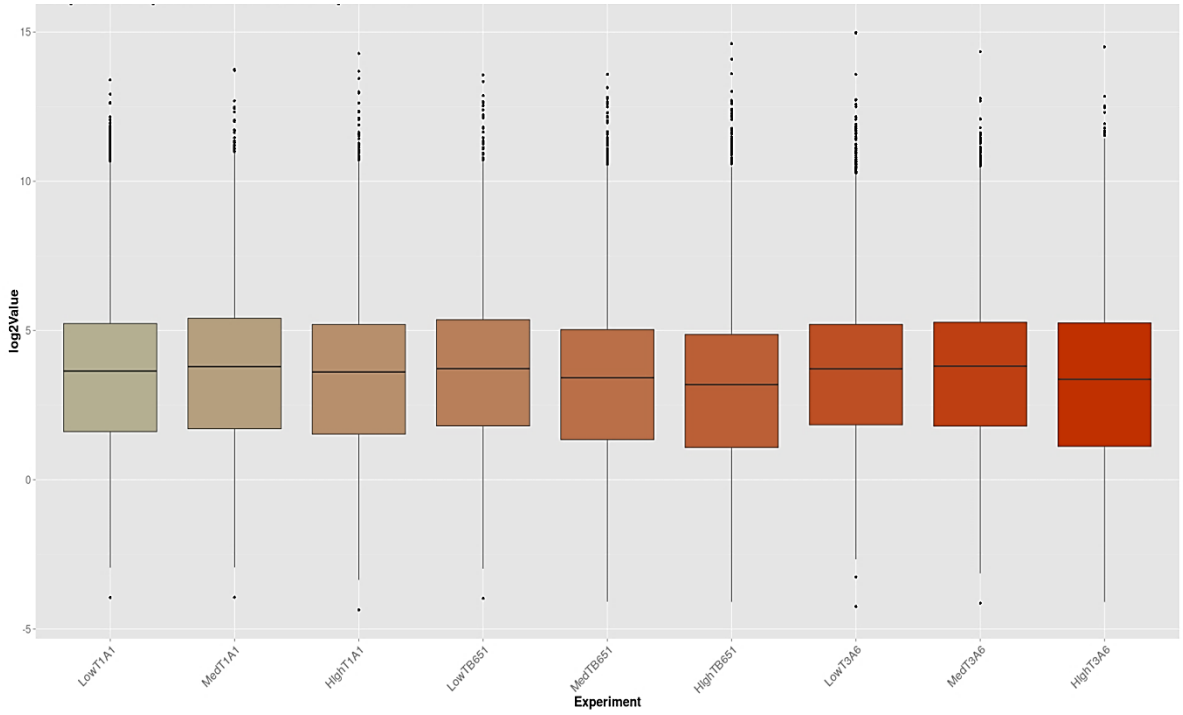
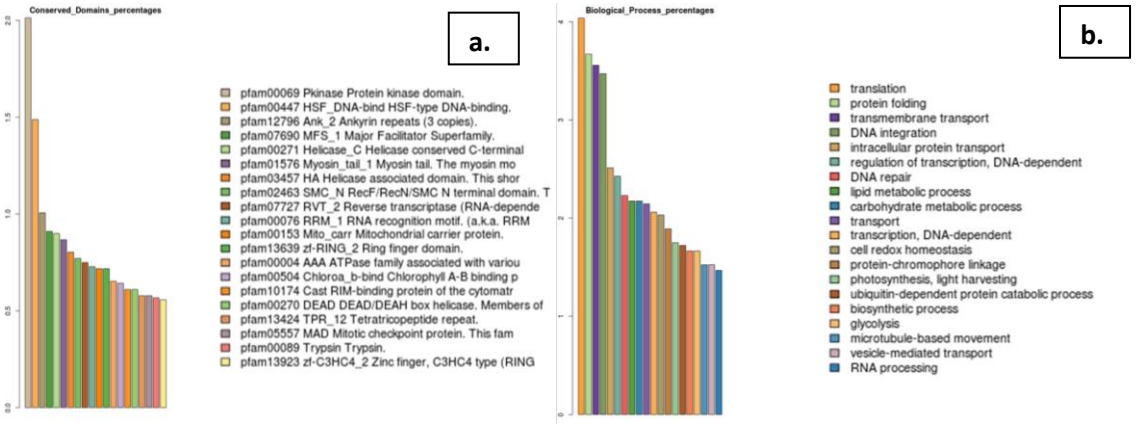


Figure 3.3.2 Boxplot of the expression values (CPMs) of each L. aporus sample (LowT : 13 °C, MedT : 19 °C, HighT : 26 °C).

The annotation of the transcriptome built based on all *L. aporus* samples showed that the two conserved domains with the higher percentage of representation in the assembly were Pkinase/Protein kinase domain (around 20%) and Heat Shock Factor (HSF)-type DNA-binding (around 15%) (Fig.3.3.3.a). The biological processes with the higher percentage were translation (more than 40%), protein folding (around 37%), transmembrane transport (36%) and DNA integration (35%) (Fig.3.3.3.b). The most represented biological process based on domains was

oxidation-reduction process (12%) (Fig.3.3.3.c) and the most represented molecular function was ATP binding (more than 15%) followed by protein binding (more than 10%) (Fig.3.3.3.d). The cellular component with the higher representation percentage was integral to membrane (more than 25%) and then nucleus (15%) (Fig. 3.3.3.e). Regarding pathways the most represented ones on level1 (lower levels represent more general categories) was protein modification (more than 20%) (Fig. 3.3.3.f), level 2 was protein ubiquitination (more than 12%) (Fig. 3.3.3.g) and at level3 pyruvate from D-glyceraldehyde 3-phosphate step (more than 20%), spermidine from putrescine step (19%), D-glyceraldehyde 3-phosphate and glycerone phosphate (17%) and pyruvate from D-glyceraldehyde 3-phosphate step (almost 15%) (Fig. 3.3.3.h).



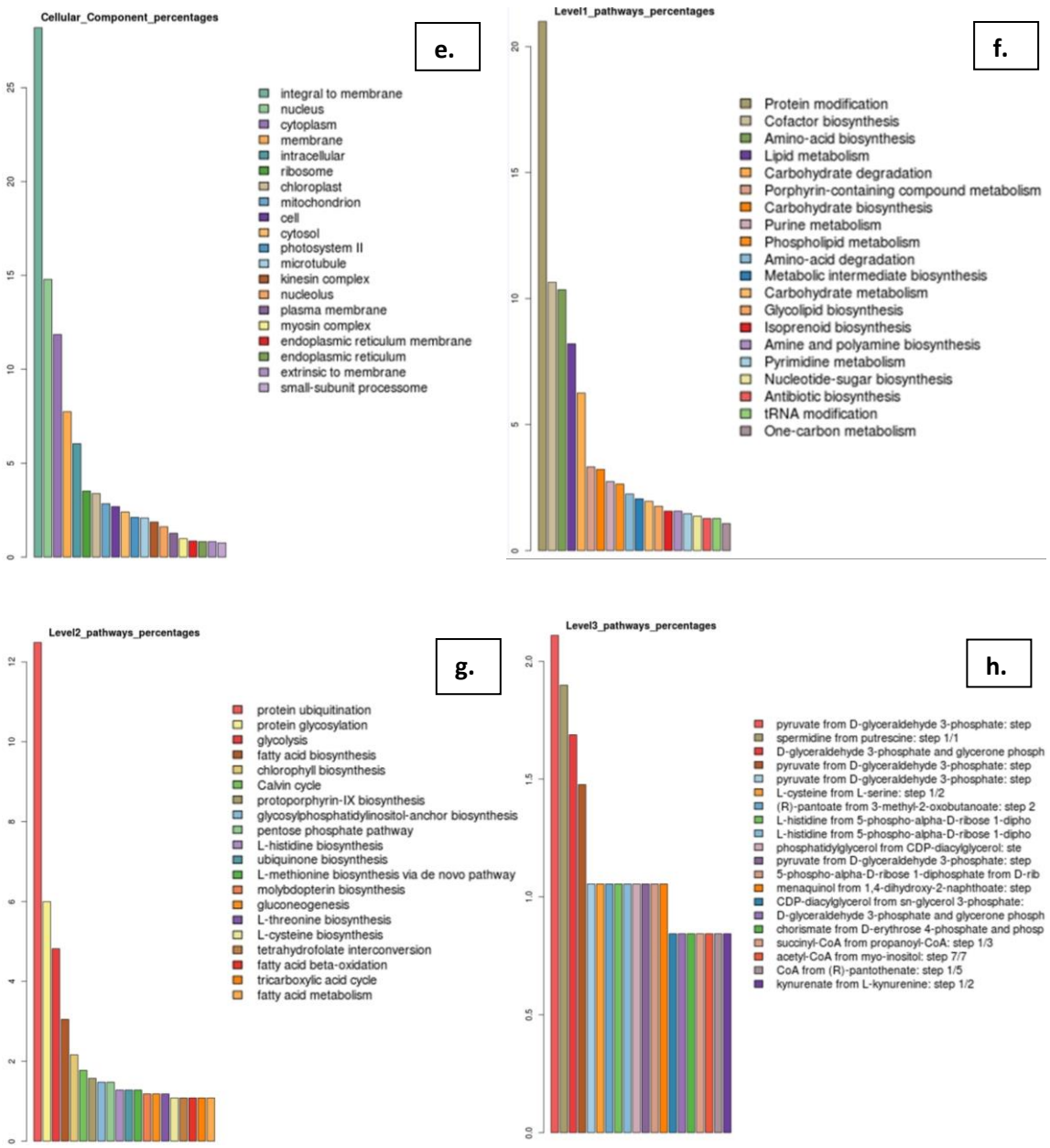


Figure 3.3.3 Bar plots of the percentages of conserved domains (a), biological process (b), biological process based on domains (c), molecular function (d), cellular component (e) GO terms, level1 (f), level2 (g) and level3 (h) pathways present in the final *L. aporus* transcriptome assembly.

The functions described above correspond to the state of *L. aporus* during the three different temperatures since the transcriptome assembly was based on reads from all available samples. So for example ATP binding and pyruvate from D-glyceraldehyde 3-phosphate step might be features of the cell under medium temperature conditions but protein folding, DNA integration and oxidation-reduction process might be result of the contribution of the low and high temperature transcripts.

3.3.1. Differential expression analysis between temperatures

The differential expression analysis between the different temperatures (using strains as replicates) resulted in 276 significantly differentially expressed (DE) genes between low (13 °C) and high (26 °C) temperature and only nine between low and medium (19 °C) temperature (Table 3.3.1.1) while there was no significant difference in expression between medium and high temperature.

Table 0.1.1 Significant DE genes between different temperatures in *L. aporus*.

	Significant DEs at low T compared to high T	Significant DEs at low T compared to medium T
Up-regulated	243	8
Down-regulated	33	1
Total	276	9

A cluster analysis of the DE transcripts shows the main response of the cells to the cold conditions (Fig. 3.3.1.1) with the exception of three clusters (clusters 4, 12 and 14).

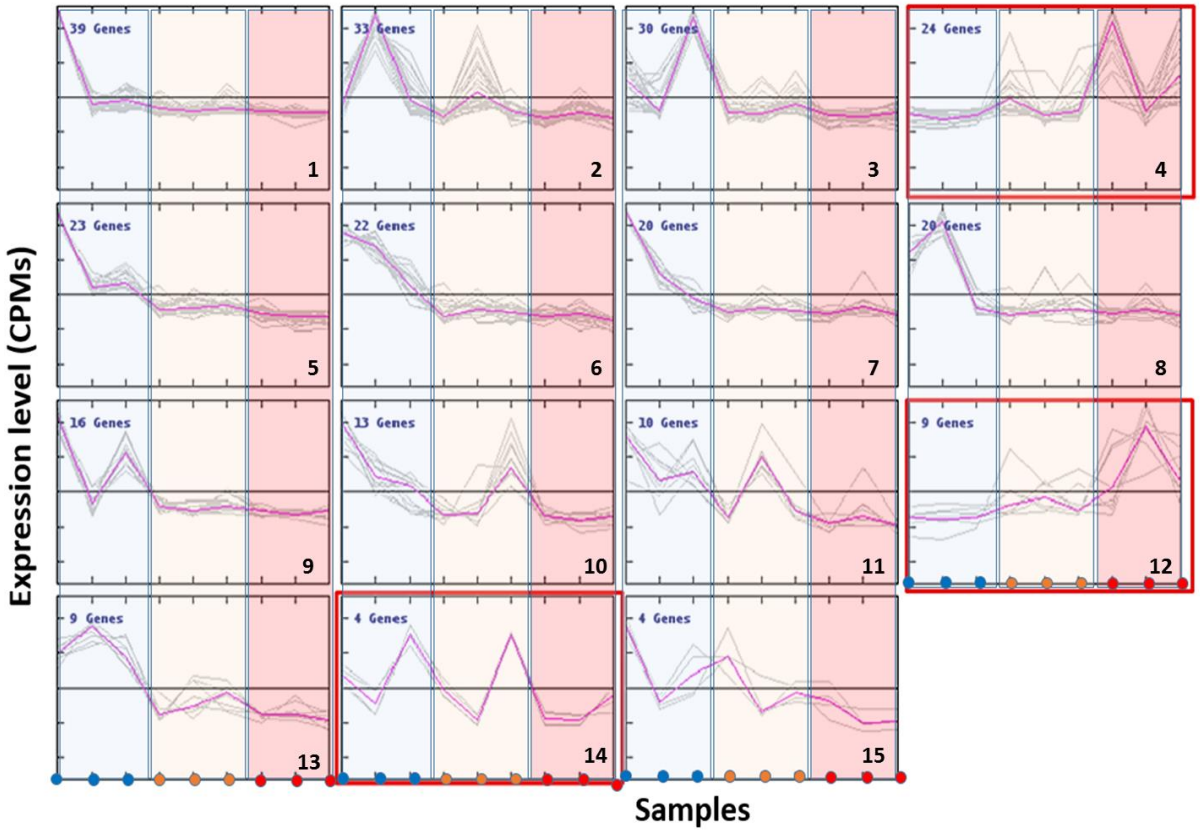


Figure 0.1.1 K-means clustering on the significant DE transcripts between low and high temperature. The dots in the x-axis correspond to the samples (B651- , 1A1-, 3A6-lowT, B651- , 1A1-, 3A6- medT, B651- , 1A1-, 3A6- highT). The low temperature samples are blue shadowed, the medium temperature ones are orange and the high temperature ones are red shadowed. Clusters are numbered on the bottom right corner while in each cluster the number of the genes included is provided on the top left corner. Clusters 4, 12 and 14 (red borderline) are deviating from the main cold responsive trend.

Several transcripts did not receive a complete annotation while the annotation of others were too generic or not considered related to the scope of this study (response to temperature and adaptation). Therefore not all significantly differentially expressed molecular functions and pathways are discussed below.

Low temperature compared to medium temperature

The most important transcripts, based on their function, among the nine differentially expressed between medium and low temperature are described below providing information reported in Annocript (Musacchia et al., 2015) and were related to:

- Antioxidant activity-oxidoreductase activity. In particular, the seed-specific plant 1-cys peroxiredoxins (PRXs) which are thiol-specific antioxidant (TSA) proteins, also known as TRX peroxidases and alkyl hydroperoxide reductase C22 (AhpC) proteins, were included. In plants, PRXs protect tissues from reactive oxygen species during desiccation; they confer a protective antioxidant role in cells through their peroxidase activity in which hydrogen peroxide, peroxynitrate and organic hydroperoxides are reduced and detoxified using reducing equivalents derived from either TRX glutathione trypanothione or AhpF. These transcripts were significantly downregulated at medium temperature.
- Methylmalonate-semialdehyde dehydrogenase acylating activity. The aldehyde dehydrogenase family (ALDH) of NAD(P) dependent enzymes in general oxidizes a wide range of endogenous and exogenous aliphatic and aromatic aldehydes to their corresponding carboxylic acids. The ALDH family plays an important role in detoxification and plant defense against oxidative stress. The *Arabidopsis* succinic semialdehyde dehydrogenase (SSADH) gene product ALDH5F1 is also included. Mutations on *Arabidopsis* ALDH5F1 result in the accumulation of H_2O_2 suggesting a role in plant defense against the environmental stress of elevated reactive oxygen species. ALDH5F1 is downregulated *L. aporus* grown at medium temperature (logFC=-4.488, FDR=0.032).

Low temperature compared to high temperature

The following graphs show the molecular functions enriched between high and low temperature (HT_LT) and are also described in detail right after.

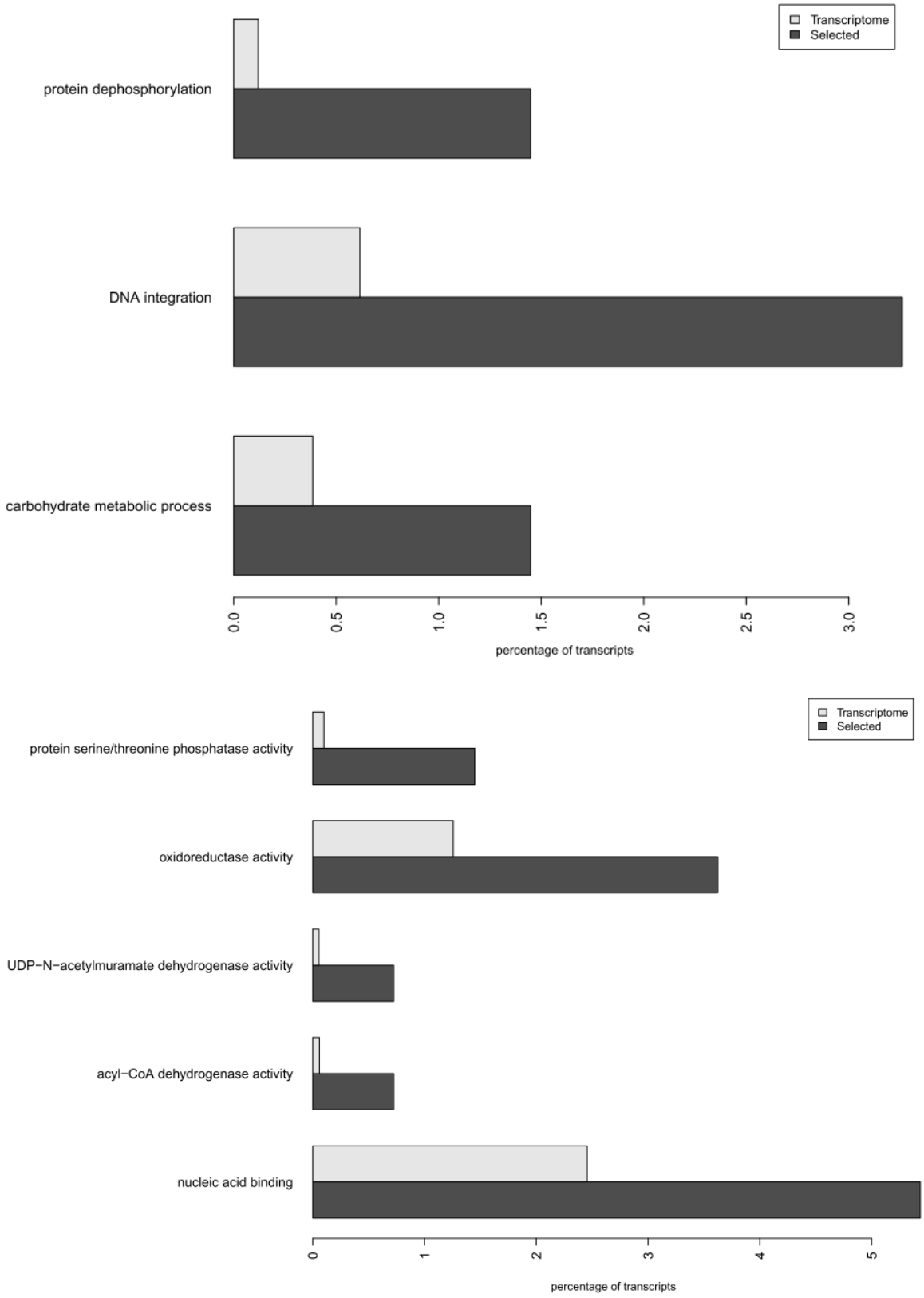


Figure 0.1.2 Biological process (above) and molecular function (below) GO terms significantly enriched in the differentially expressed genes between high and low temperature. Selected refers to the significantly differentially expressed transcripts and transcriptome refers to the total *L. aporus* transcripts.

In more details, the enriched processes and functions that were considered more interesting due to a possible link with the low temperature response were:

- Protein dephosphorylation and protein serine/threonine phosphatase activity. The related transcripts code for serine/threonine-protein kinase SAPK2 or in CDD description STKc_AMPK-like catalytic domain of AMP-activated protein kinase-like serine/threonine kinases and probable serine/threonine-protein kinase fhkD. They serve as master regulators of glucose and lipid metabolism by monitoring carbon and energy supplies via sensing the AMP/ATP ratio of the cell. The large family of serine/threonine kinases (STKs) regulates many cellular processes including proliferation, division, survival, metabolism and cell-cycle progression. SAPK2 is downregulated at low temperature *L. aporus* (logFC=-4.612, FDR=0.0009) and fhkD is upregulated (logFC=3.801, FDR=0.022).
- DNA integration refers to the process in which a segment of DNA is incorporated into another, usually larger, DNA molecule such as a chromosome. This process includes transposon integration. Nine transcripts were related to DNA integration.
- Oxidoreductase activity. This function includes the UDP-N acetylmuramate dehydrogenase activity which is also enriched at low temperature. The transcripts involved in this activity were mainly related to redox sensing such as transcripts encoding for ferredoxin-NADP reductase and peroxiredoxin. In addition there were transcripts encoding for proteins belonging to the short-chain dehydrogenase/reductase (SDR) superfamily such as the light-dependent protochlorophyllide (Pchlide) oxidoreductase (LPOR). LPOR is one of the two unrelated Pchlide reductases in the penultimate step of chlorophyll biosynthesis; the other one is dark-operative Pchlide reductase (DPOR), a nitrogenase-like enzyme sensitive to oxygen.
- Nucleic acid binding. This molecular function appears to be one of the most enriched ones among the significantly differentially expressed transcripts. When checking these transcripts into more details it is concluded that all of them (except a few that have no detailed annotation) are also associated with DNA integration.

Amino-acid degradation was the most enriched pathway with respect to the overall transcriptomes, indicating failed metabolic compensation (Fig. 3.3.1.3). This pathway includes a diverse set of enzymes many of which play important roles in fatty acid metabolism, like the crotonase/Enoyl-coenzyme A (CoA) hydratase superfamily. The transcripts involved were upregulated at low temperature.

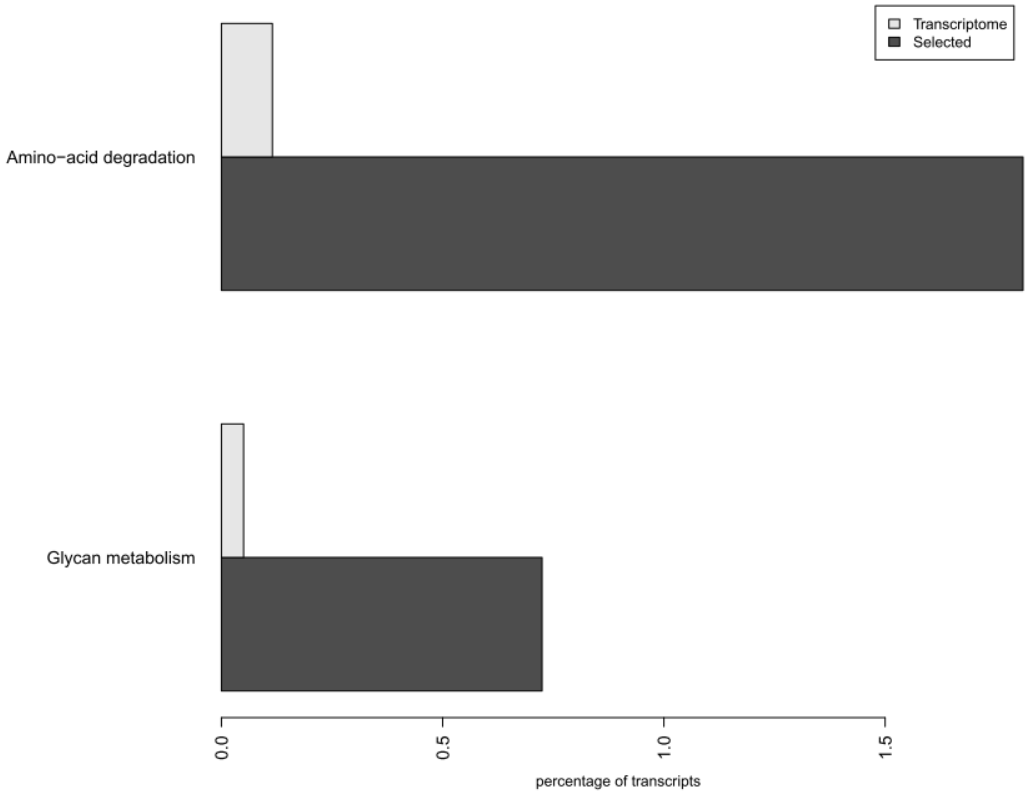


Figure 0.1.3 Pathways enriched in *L. aporus* differential expressed genes between high and low temperature. Selected refers to the significantly differentially expressed transcripts and transcriptome refers to the total *L. aporus* transcripts.

Transcripts related to specific functions and conditions such as temperature, response to stress, nutrients, adaptation, growth, sexual reproduction and light were searched for among the significantly differentially expressed transcripts. There were 13 transcripts related to heat stress, seven related to transposable elements, one temperature related and seven environmental stress-related (in addition to the ones mentioned in the enrichment analysis):

- ✓ Response to thermal stress/ Heat shock response and temperature related transcripts. In this group i) eight chaperones related transcripts are included since many chaperones are heat shock proteins (HSPs), ii) four transcripts of heat shock factor proteins and heat

stress transcription factors, iii) one transcript of photosystem II 12 kDa extrinsic protein (PSBU), a protein that stabilizes the structure of photosystem II oxygen-evolving complex (OEC), the ion environment of oxygen evolution and protects the OEC against heat-induced inactivation, and iv) one transcript of SYM1, a stress-induced protein related to temperature changes. They are all significantly upregulated at low temperature except for heat shock factor protein 1 (HSF1) and heat stress transcription factor C-1b (HSFC1B).

- ✓ Transposable elements. The transcripts found in addition to the DNA integration described above were transcripts annotated as domains of reverse transcriptase, ribonuclease, transposase (domains commonly met in transposons) or as insertion element.
- ✓ Environmental stress. (i) Alternative oxidase AOX4 which is a ubiquinol oxidase acting early in chloroplast biogenesis as a component of a redox chain responsible for phytoene desaturation in plants, some fungi and protists. It prevents the generation of toxic oxygen radicals and photooxidation of the nascent photosynthetic apparatus. In our dataset AOX4 was found significantly upregulated in low temperature *L. aporus* (logFC=4.657, FDR=0.0005). (ii) Methylmalonate-semialdehyde dehydrogenase (Aldh6a1), an enzyme involved in valine and pyrimidine metabolism. The ALDH family was already described above and it is upregulated in low temperature *L. aporus* (logFC=3.541, FDR=0.035) compared to high temperature. (iii) Ribosome associated inhibitor A (RaiA), also known as Protein Y (PY), YfiA and SpotY, is a stress-response protein that binds the ribosomal subunit interface and arrests translation by interfering with aminoacyl-tRNA binding to the ribosomal A site. RaiA is also thought to counteract miscoding at the A site thus reducing translation errors. The RaiA fold structurally resembles the double-stranded RNA-binding domain (dsRBD). The same transcript includes domains related to the sigma-54 modulation protein family and the S30AE family of ribosomal proteins which includes the light- repressed protein (lrtA). It is upregulated at low temperature. (iv) Formate dehydrogenase like. Formate dehydrogenase (FDH) catalyzes the NAD+-dependent

oxidation of formate ion to carbon dioxide with the concomitant reduction of NAD+ to NADH. FDHs are found in all methylotrophic microorganisms in energy production and in the stress responses of plants. It is part of two transcripts that are both upregulated at low temperature. (v) ABC transporter periplasmic binding protein YdcS which is involved in uptake of polyamines. Polyamine transport plays an essential role in the regulation of intracellular polyamine levels which are known to be elevated in rapidly proliferating cells and tumors. Natural polyamines are putrescine, spermidine, and spermine. They are polycations that play multiple roles in cell growth, survival and proliferation, plant stress and disease resistance. They can interact with negatively charged molecules, such as nucleic acids, to modulate their functions. The two related transcripts are upregulated at low temperature.

Uncharacterized proteins

TR6937|c1_g1_i1. This transcript based on blastx in NCBI could translate to a conserved hypothetical protein found in the bacterium *Magnetospirillum gryphiswaldense* (88% identity, 83% coverage, evalue=2e-36). Magnetosome island (MAI) is related to biomineralization, contains numerous genes with unknown functions and it is rich in insertion elements. This transcript is overexpressed at low temperature.

TR4495|c0_g1_i1. This transcript contains a conserved domain of 2Fe-2S ferredoxin (iron-sulphur protein) binding site. 2Fe-2S ferredoxins play an important role in electron transfer processes such as in photosynthesis. Several oxidoreductases contain redox domains similar to 2Fe-2S ferredoxins implying that this low T-upregulated transcript could be included in the group related to oxidative stress.

TR7159|c1_g1_i1. A conserved domain of this transcript belongs to transcription factor PAP1 which regulates antioxidant-gene transcription in response to H_2O_2.

TR6538|c6_g1_i2. Possible transcript of the hypothetical protein mgI388 belonging to the MAI region mentioned above (71% identity, 82% coverage, evalue=7e-20). It is upregulated at low temperature as well.

TR5929|c0_g1_i1. Possible transcript of the hypothetical protein mgI382 belonging to the MAI region mentioned above (72% identity, 95% coverage, 1e-27). It is upregulated at low temperature as well.

3.3.2. Differential expression analysis between strains

Screening the expression values of the interesting genes mentioned in section 3.3.1., a strain specific pattern was noted. In order to confirm this, we proceeded with a PCA analysis of all samples and a heatmap of the transcripts that showed highly differential expression in at least one of an in-between-strains comparison made in T-REx (HighFold: fold change ≥ 5 and p-value ≤ 0.01) (Fig. 3.3.2.1. and 3.3.2.2).

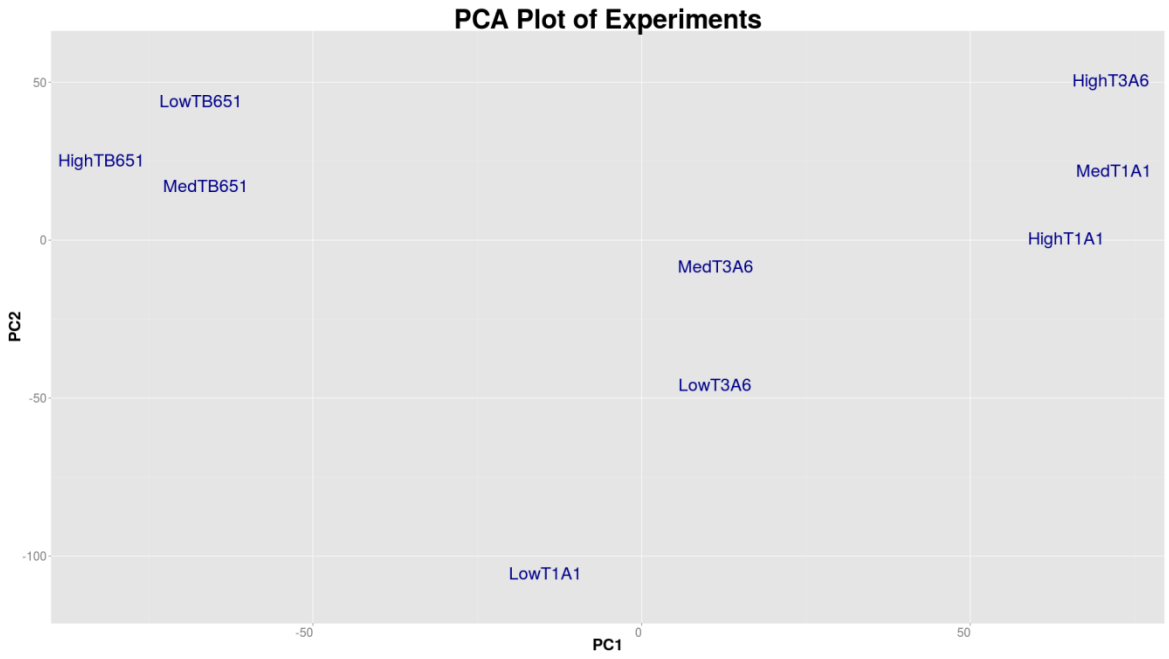


Figure 0.2.1 PCA analysis of all *L. aporus* samples based on the expression values (CPMs) of all transcripts. LowT: low temperature, MedT: medium temperature, HighT: high temperature.

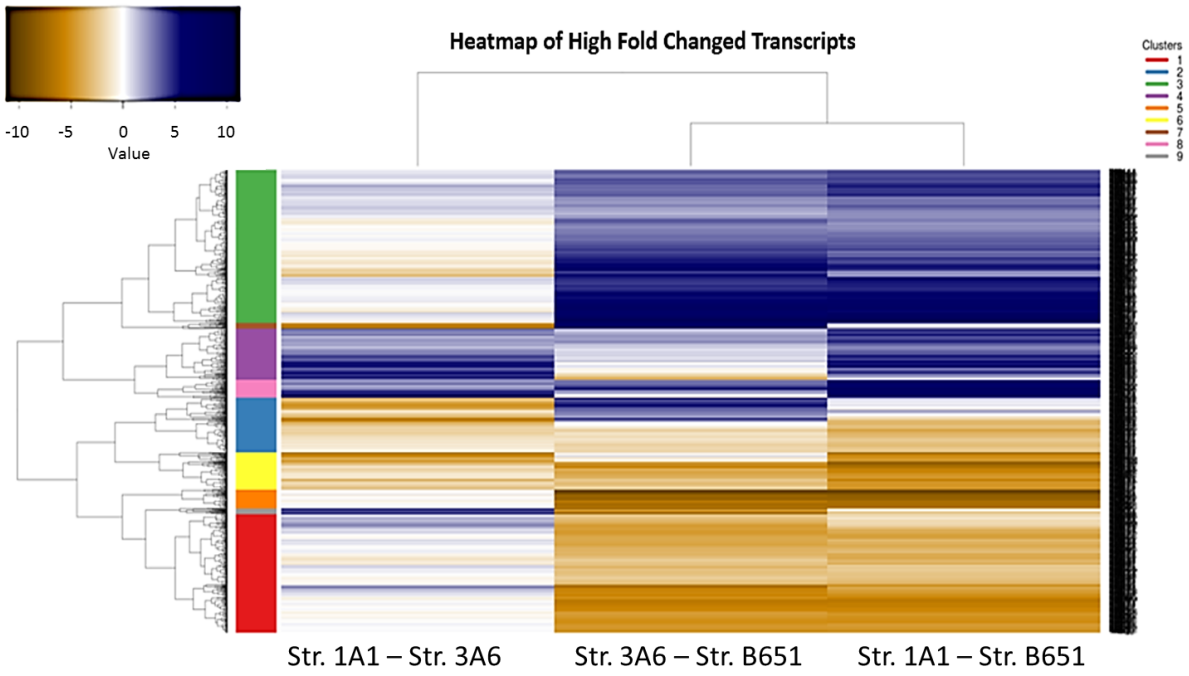


Figure 0.2.2 Heatmap and corresponding clustering based on expression values (CPMs) of the high fold changed transcripts in the between *L. aporus* strains comparison done in T-REx.

In the PCA of the expression values of all samples, B651 seems to cluster apart from the other two strains that group based on temperature rather than strain. At the same time the heatmap of the high fold transcripts shows clearly a much lower difference in the expression patterns of the 1A1-3A6 pair compared to the pattern when B651 is included in the pair, thus proving the higher resemblance of 1A1 and 3A6. Based on these evidences, a second differential expression analysis (following the exact same procedure as the first analysis between temperatures) was performed but this time between strains (using the different temperatures as replicates). 622 transcripts were significantly differentially expressed between 1A1 and 3A6; much less compared to the 3,015 between 1A1 and B651 and 2,418 between B651 and 3A6 (Fig. 3.3.2.3).

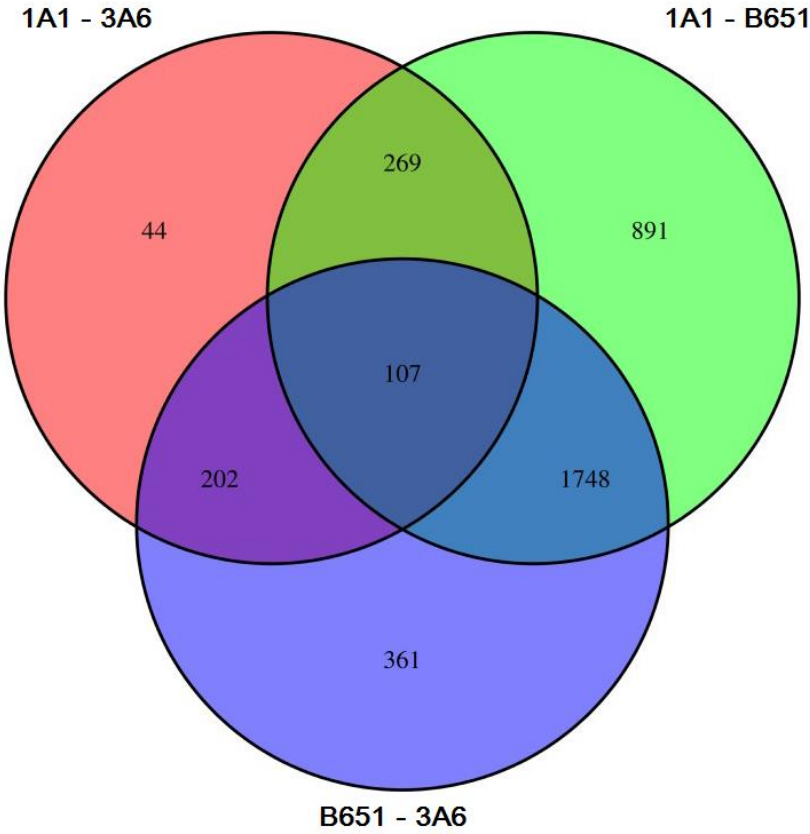


Figure 0.2.3 Venn diagram of the significantly differential expressed genes between strains in *L. aporus* when temperature conditions are used as replicates.

Transcripts with a fold change larger than 10 or smaller than -10 were considered as highly expressed; 35 transcripts were highly expressed in the 3A6 – 1A1 pair, 187 in the 1A1 – B651 pair and 142 in the B651 – 3A6 showing again a higher difference when B651 is included in the pair. The enrichment analysis resulted in a considerable number of GO terms related to DNA integration: 20.33% (25/123) in the 1A1 - 3A6 pair, and double or more in 1A1 - B651 (49.59%, 61/123) and in B651 – 3A6 (54.47%, 67/123). DNA integration was the biological process enriched in all three pairs so the TE-related transcripts detected in the temperature DE analysis were searched for and **all were found** among the significantly differentially expressed ones of this analysis as well. The pathway enrichment analysis had no results on the 1A1 -3A6 pair while isoprenoid biosynthesis was the pathway enriched in 1A1 – B651 and antibiotic biosynthesis in the B651 – 3A6 pair.

3.3.3. Transposon-related analysis

Most of the DNA-integration transcripts were expressed in a similar manner, i.e., lower at low temperature and lower in B651; there were even some transcripts completely absent from all B651 samples. Therefore a further analysis on the transposon related transcripts was designed here. The transcripts that were found significantly differentially expressed between temperatures were searched for an expression less than a half in B651 compared to the other two strains. 83 out of the 277 significantly differentially expressed transcripts were found to be expressed at a low level in B651 compared to the other two strains. All sixteen DNA integration/transposon related transcripts were present in the list. Hierarchical clustering was performed in MeV for these TE transcripts (Fig. 3.3.3.1).

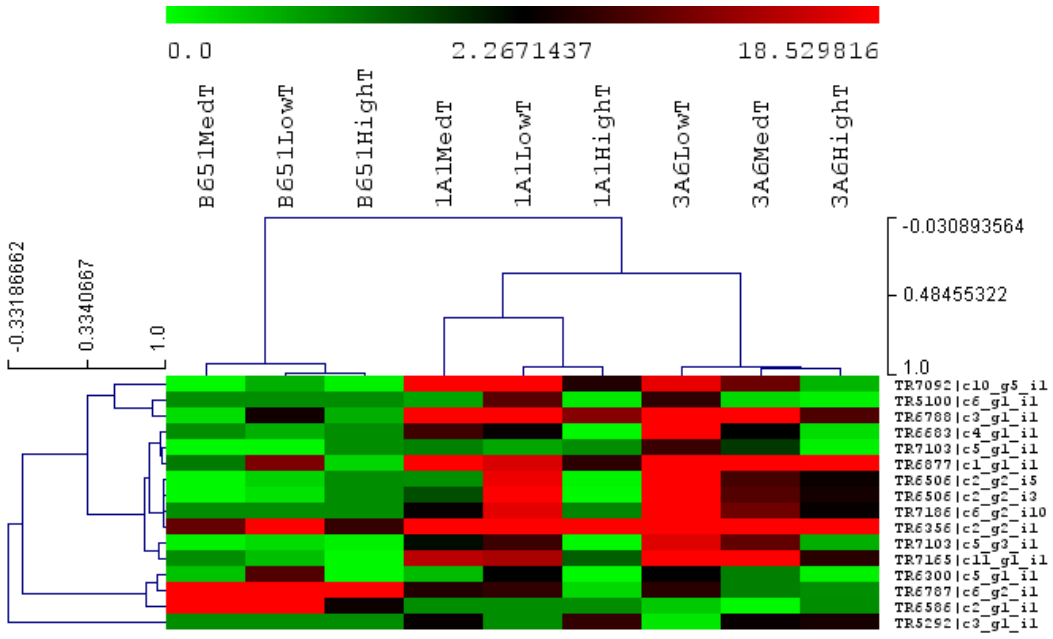


Figure 0.3.1 Hierarchical clustering and corresponding heatmap of all DE transcripts related to TEs, found significant both between temperatures and between strains in *L. aporus*.

On the other hand, the stress and heat/temperature related transcripts were all absent from the list except for three, AOX4, HSFC1b and HSFA1a. The clusters produced in section 3.3.1 (Fig. 3.3.1.1) were searched for the TE related transcripts as well. Cluster 3 included four TE related transcripts and HSFA1a (Fig. 3.3.3.2).

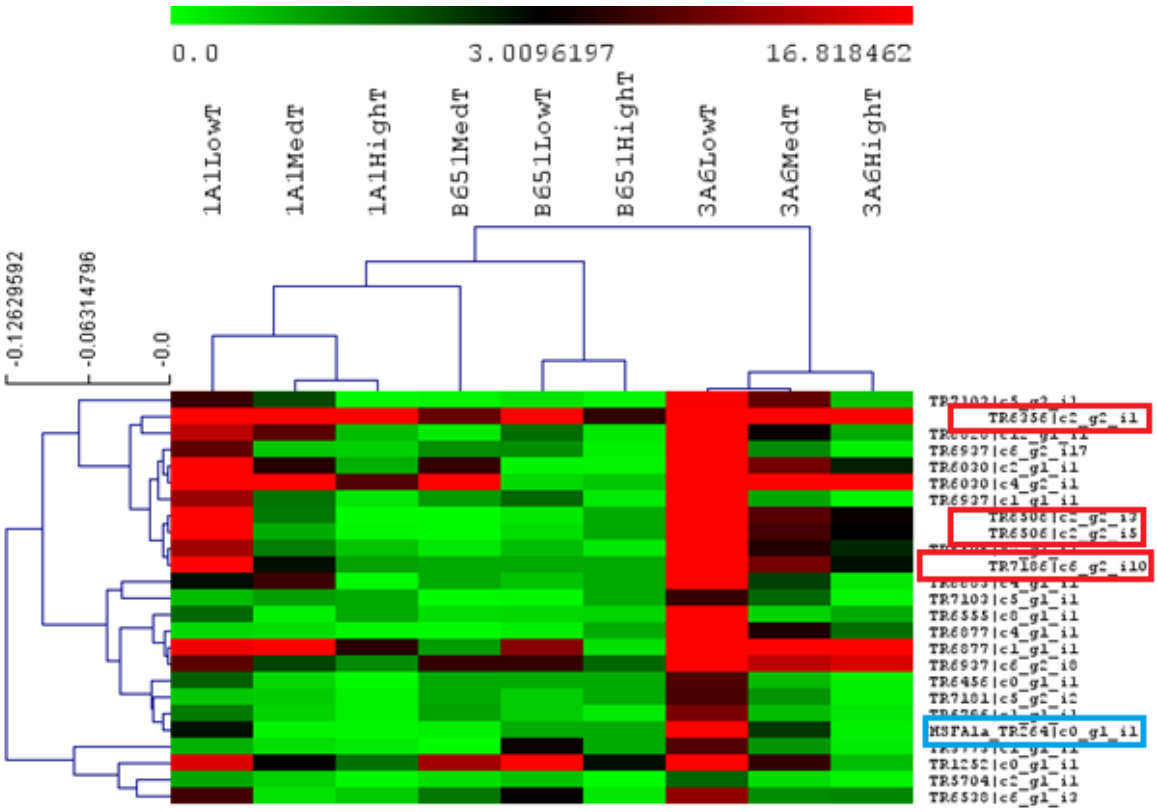


Figure 0.3.2 Heatmap of cluster 3 produced by the k-means clustering in the DE analysis between temperatures in section 3.3.1. TE-transposons are highlighted red and HSFA1a is blue.

Based on the blat results in Ensembl, eleven transposon-related transcripts were grouped and then used to build a phylogenetic tree. Four main groups were created; group A, B and group C belong to the Ty1 Copia transposon family while group D is a Ty3 Gypsy transposon family. Both types of TEs are included in the LTR order of retrotransposons. Group A and B transposons overlap mainly on the reverse transcriptase domain of the corresponding proteins while group C on the ribonuclease H like domain. Group D transposons overlap on the ribonuclease H-like domain, reverse transcriptase and integrase domain. The expression values of all groups can be seen in Fig. 3.3.3.3. Group B includes only one significant DE transcript while the other two are possible isoforms.

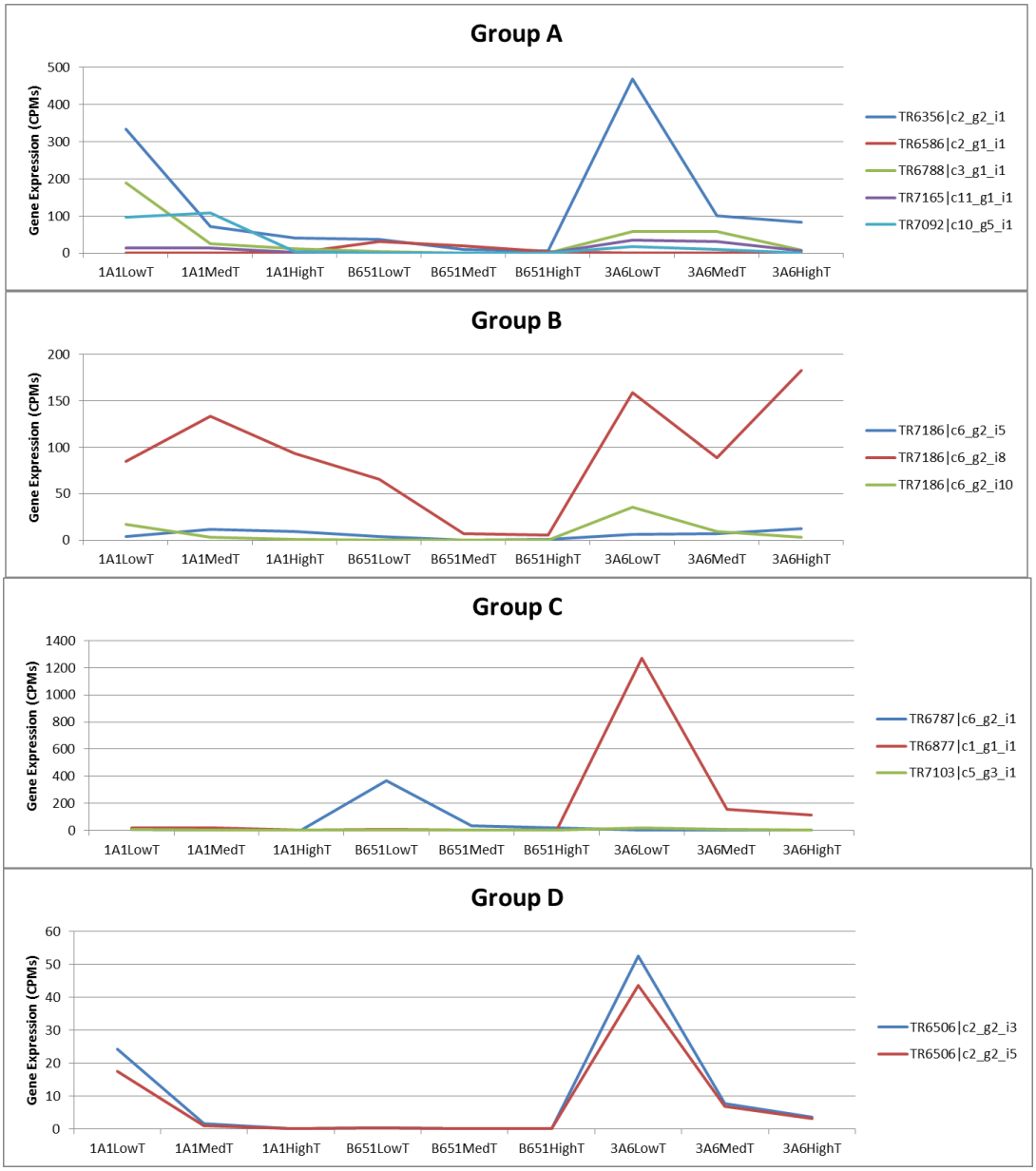


Figure 0.3.3 Expression values (CPMs) in all *L. aporus* samples of the groups of TE related transcripts. Group B includes only one significant DE transcript (TR7186|c6_g2_i10) while the other two are possible isoforms.

A phylogenetic tree was constructed only for the DEs of group A and B because they were the only groups that shared common sites (Fig. 3.3.3.4). The selected transcripts are from different clades.

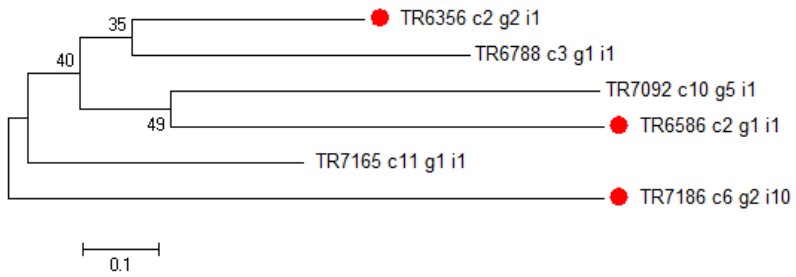


Figure 3.0.4 ML tree with 500 bootstrap, Poisson model and pairwise deletion of group A and group B transposons. The selected transcripts for validation and further qRT-PCR experiments are highlighted in red.

The Ty3_Gypsy group (group D) only consisted of two transcripts, which were also selected for validation. No tree was constructed for this group either but the two sequences were found to be 96% identical in the nucleotide level (95% coverage) and 97% in the amino acid level (84% coverage).

When screened in RepeatMasker, three out of the sixteen TE related transcripts matched with CoDi transposons. TR7186 matched with Blackbeard, TR6877 matched with CoDi6 and TR6506_i3 matched with GyDi2 which is Gypsy-like element.

In the end, five TE related transcripts that match the hypothesis of the stress-induced transposons were selected: **TR6356, TR7186, TR6586, TR6506_i3** and **TR6506_i5** (Table 3.3.3.1 and 3.3.3.2). The differentiation of B651 was also taken into account. In particular, the final selection of the TE transcripts to be further investigated was based on the following criteria:

- The transcripts selected were all less expressed at low temperature and even less in the B651 samples except TR6586 which showed the same relationship with low temperature but not in B651 (higher values in B651 than in 1A1 and 3A6). TR6586 was also included due to its interesting reverse strain expression compared to the others.
- All transcripts were included in cluster 3, except for TR6586 (Fig.3.3.3.2).
- Group A and B transcripts were preferred over group C because of their expression patterns (low expression at 13 °C and in B651) which were more suitable to our hypothesis (Fig. 3.3.3.3).
- Transcripts were selected from different phylogenetic clades (Fig. 3.3.3.4). TR6506_i3 and _i5 were the only ones from the Ty3/Gypsy family.
- TR7186 was included also due to tis similarity to the Blackbeard transposon which has already been discussed in the introduction.

Table 0.3.1 Selected TE related transcripts and their corresponding RNA-seq expression values (CPM values) for each sample. Blue columns correspond to 13°C, yellow to 19°C and red to 26°C.

ID	B651			IA1			3A6		
TR6356	36,227	8,734	5,693	333,9	71,383	41,206	468,23	100,02	83,154
TR7186	0	0	0	16,958	2,811	1,075	35,427	9,388	2,811
TR6586	31,468	18,53	3,169	0	0	0	0,473	0,057	0
TR6506_i3	0,254	0,059	0	24,3	1,569	0,049	52,458	7,68	3,631
TR6506_i5	0,381	0,059	0	17,608	0,98	0,098	43,522	6,77	3,045

Table 0.3.2 Selected TE related transcripts and their corresponding encoded domains

Transcript_ID	Gene/Domain_Name to be Amplified
TR6356|c2_g2_i1	RNase_HI_RT_Ty1/Copia family
TR7186|c6_g2_i10	RNase_HI_RT_Ty1/Copia family
TR6586|c2_g1_i1	RVT_2 Reverse transcriptase
TR6506|c2_g2_i3	rve Integrase core domain (Ty3/Gypsy family
TR6506|c2_g2_i5	rve Integrase core domain (Ty3 /Gypsy family)

3.3.4. qRT-PCR analysis

For the qRT-PCR validation and experiment, three more genes were added to the TE related ones that were selected in section 3.3.3. The three extra transcripts were all related to temperature and they were: heat transcription factor A-1a (**HSFA**), heat stress transcription factor B-2a (**HSFB**) and protein **SYM1**. The temperature related transcripts were added as a backup check for the cold stress. They were selected based on the available bibliography on their relation to heat/ cold stress and especially HSFA1a had to be included in order to further investigate a possible co-regulation with the TEs (co-expressed in cluster 3, Fig. 3.3.3.2). As most significant differences in the differential expression analysis among temperatures were seen between the highest and the lowest temperature, the calculations were performed between these two temperatures.

Validation Set

The qRT-PCR results are presented along with the RNA sequencing results for each transcript and for each sample, expressed as logarithmic fold change of low to high temperature (Fig. 3.3.4.1). A log2FC ≥ 1 means that the gene is at least two times higher expressed at low temperature compared to high temperature (FC ≥ 2). Accordingly a log2FC ≤ -1 means that the gene is at least two-fold downregulated at low temperature compared to high. All the transcripts of the “2015”

samples were upregulated at low temperature following the tendency of "r" samples (RNA-seq samples) therefore validating the RNA-seq results with only a couple of exceptions in the temperature related transcripts (HSFA in B651 2015 and HSFB in 1A1 2015).

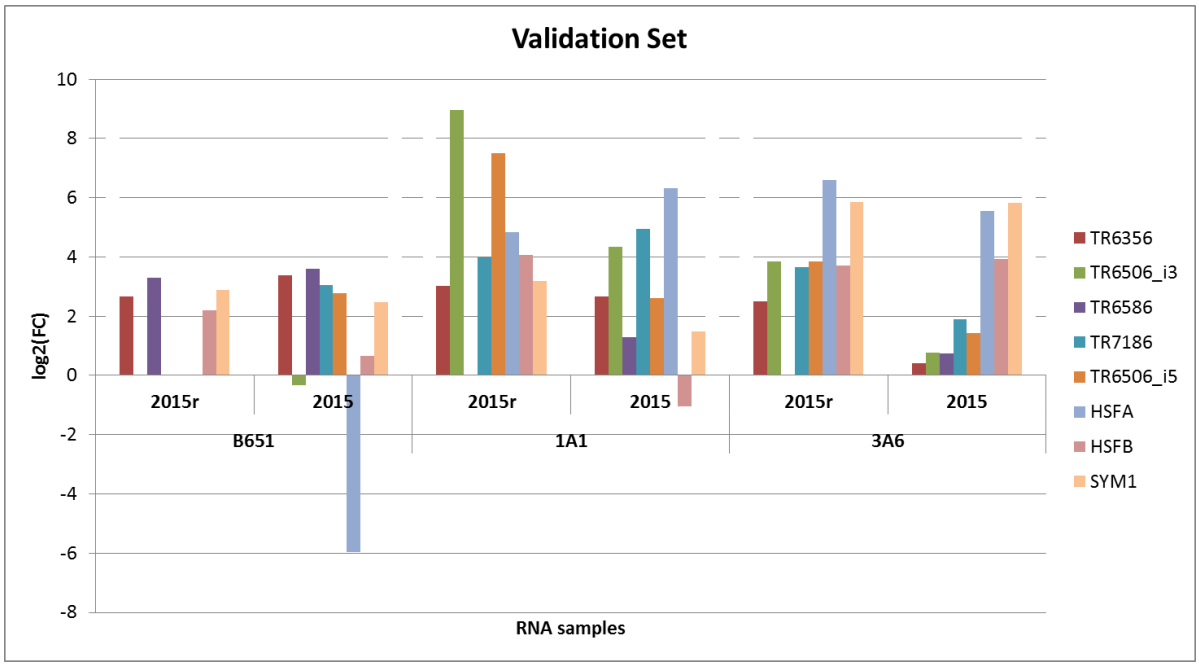


Figure 0.4.1 qRT-PCR results for TE and temperature related transcripts presented together with RNA sequencing results for each sample of the validation set. The fold change is low temperature to high temperature expression values. 2015r: RNA sequencing samples, 2015: qRT-PCR results from 2015 acclimatized/ validation set samples.

In this graph, B651 stands out based on its different gene expression level compared to 1A1 and 3A6, especially in the case of TR6506_i3, TR6586 and HSFA.

Exploration Set

While the validation of the RNA seq results was successful, the qRT-PCR experiments performed with the same strains acclimatized in other periods or with newly isolated strains showed quite different results. In fact only SYM1 was upregulated at the low temperature compared to the high temperature while several other genes were downregulated (Fig. 3.3.4.2).

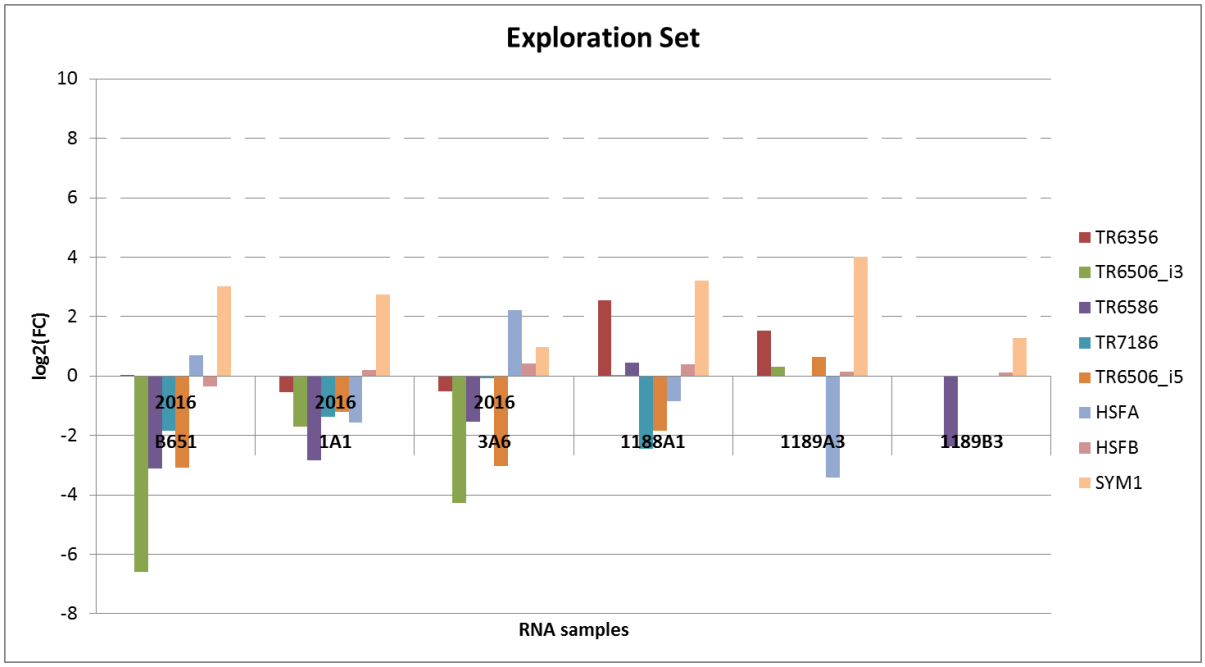


Figure 0.4.2 qRT-PCR results for TE and temperature related transcripts for each sample of the exploration set. The fold change is low temperature to high temperature expression values. 2016: qRT-PCR results from 2016 acclimatized/ exploration set 1 samples, 1188A1; 1189A3; 1189B3: exploration set 2 samples.

The disagreement in the expression of the TE related transcripts of the "2016" qRT-PCR results and the RNA-seq results were assumed to be linked with the difference in the acclimatization period. The samples used for RNA-seq and for the "2015" qRT-PCR validation came from strains acclimatized for a longer period compared to the "2016" samples. In order to confirm this assumption the growth rates of the strains during these two periods were compared.

The growth responses of the various strains during the last three growth cycles before filtration and RNA extraction are shown in Fig. 3.3.4.3. The samples grown for RNA sequencing are included for a comparison. A slight difference between 2015 and 2016 samples can already be observed.

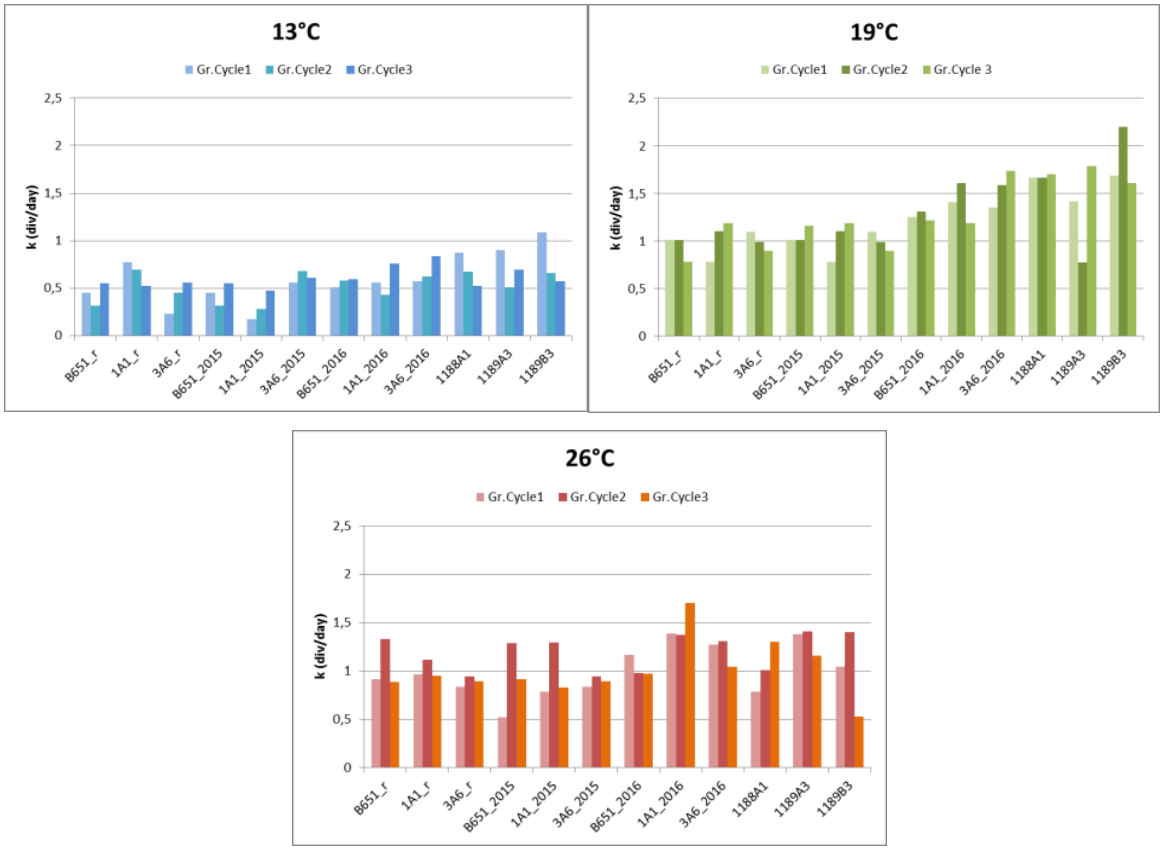


Figure 0.4.3 Bar graph of growth rates of *L. aporus* strains at the three different growth temperatures, based on fluorescence. _r: RNA sequencing samples, _2015: validation set samples acclimatized in 2015 for the same time as _r samples, _2016: exploration set 1 samples acclimatized in 2016, 1188A1; 1189A3; 1189B3: exploration set 2 samples.

The possible cause for this difference could be (i) the difference in duration of acclimatization periods of samples as already mentioned in the assumption, (ii) the different seasons of isolation and/or (iii) the different season when the experiments were conducted. Therefore statistical tests were performed for each case. When values were normally distributed and homogeneous (Kolmogorov-Smirnov and Levene statistic test respectively, $p>0.05$) one-way ANOVA was used with a Tukey post hoc test. In any other case, a Welch and a Brown-Forsythe ANOVA were performed with Games-Howell post hoc test.

Overall, the statistical tests performed pointed towards a dependence of the growth rate on the acclimatization time and the filtration dates (Table 3.3.4.1).

Table 0.4.1 Summary of the statistical tests done on the growth rates of the strains acclimatized for RNA-seq and qRT-PCR experiments.

Statistical Tested Growth Rates (13 °C, 19 °C, 26 °C separately)	Statistical Test	Post hoc test
1A1 2015r, 3A6 2015r, B651 2015r 1A1 2015, 3A6 2015, B651 2015 1A1 2016, 3A6 2016, B651 2016	13 °C, 19 °C, 26°C: one-way ANOVA, $p < 0,05$	13 °C: 1A1 2015 - 3A6 2016, 1A1 2015 - 1A1 2015r 19 °C: B651 2015r - 1A1 2016, 3A6 2016 - all except for 1A1 and B651 2016 26 °C: 1A1 2016 - 3A6 2015r, 1A1 2016 - B651 2015, 1A1 2016 - 3A6 2015
1A1 2016, 3A6 2016, B651 2016 1188A1, 1189A3, 1189B3	13 °C, 26 °C: one-way ANOVA, $p > 0,05$ 19 °C: Brown-Forsythe ANOVA, $p > 0,05$ and Welch ANOVA, $p < 0,05$	19 °C: B651 2016 - 1188A1
Isolation dates of all	13 °C, 19 °C: one-way ANOVA, $p < 0,05$ 26 °C: one-way ANOVA, $p > 0,05$	13 °C: February - August 19 °C: December - February, January - February, February - August
Measurement and filtration dates of all	13 °C, 19 °C: one-way ANOVA, $p < 0,05$ 26 °C: Brown-Forsythe and Welch ANOVA, $p > 0,05$	13 °C: January - March, January - May, February - March, February - April, February May 19 °C: December - March, December - April, February - March, February - April
Measurement and filtration dates of “2015” samples	13 °C, 19 °C: one-way ANOVA, $p < 0,05$ 26 °C: Brown-Forsythe and Welch ANOVA, $p < 0,05$	13 °C: March - January, March - February 19 °C: March - December, March - February 26 °C: March - January

The significant difference of growth rates mainly related to February as the isolation month is not so straightforward considering that February was actually the month in which all 2016 samples with the shortest acclimatization period were isolated. Based on that, the isolation date effect was rejected. On the other hand, despite the separation of the growth rates of samples measured and filtered in winter – spring, it should be again noted that all the May and April samples came from 2016 experiments. In order to homogenize the dataset, the statistical test on the filtration dates was performed again only on the 2015 samples. The 2016 samples were not tested since they were all spring measurements.

For some transcripts the expression level was too low or there was no expression detected at all in specific samples, mainly the newly isolated or else short-acclimatized ones (Table 3.3.4.2).

Table 0.4.2 Expression of selected transcripts in samples used for qRT-PCR. Green: present, red: absent (samples with ct>ct^{negative} were as well considered absent), orange: very low expression (ct>=34< ct^{negative}), bold crosses: validated for RNA-seq, asterisks: validated also for B651 differential expression. Ct is the cycle threshold which is defined as the number of PCR cycles required for the signal of the product to exceed the background level.

Strain	B651			1A1			3A6			1188A1			1189A3			1189B3		
T (°C)	13	19	26	13	19	26	13	19	26	13	19	26	13	19	26	13	19	26
6356	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	-	-	-
*6506_i3	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	-	-	+
6506_i5	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	-	+	+
*6586	+	+	+	+	+	+	+	+	+	+	+	+	-	-	-	-	-	+
*7186	+	+	+	+	+	+	+	+	+	+	+	+	-	-	-	-	-	-
HSFA	+	+	+	+	+	+	+	+	+	+	+	+	+	-	+	-	-	-
HSFB	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
SYM1	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+

It should be noted that TR6356, TR6586 and TRi5 primers were not equally efficient for all samples. In particular, double or different melting curves were noticed in 1A1, 1188A1, 1189A3 and 1189B3, especially in 13 °C, implying no specific amplification in these samples. This phenomenon was also obvious in 3A6 but in a much lesser degree while in B651 was completely absent (one clear peak in all conditions). The primer dimer peak was especially high for HSFA in the samples 1A1 19 °C 2016 and 1A1 26 °C 2015. Because the background is higher, large amounts of primer dimers may alter the Ct of the experimental samples and change the expression level interpretation (In qRT-PCR, the progress of the PCR reaction is measured by accumulation of a fluorescence signal during amplification; Ct is the cycle threshold which is defined as the number of PCR cycles required for the fluorescent signal to cross the threshold i.e. exceeds background level). Therefore the final expression of these samples could be actually lower than the one depicted in the graph. This might be an effect of the much lower level of these transcripts in these strains which leads to an amplification of a different product or formation of primer dimers.

3.3.5. Search for the validated transcripts in other datasets

The transcripts that were chosen to be validated with qRT-PCR, as well as the transcripts of heat shock factor protein 1 (TR6847|c0_g1_i1), heat stress transcription factor C-1b (TR1078|c1_g1_i1), Photosystem II 12 kDa extrinsic protein chloroplastic (TR268|c0_g1_i1) and ubiquinol oxidase 4 chloroplastic/chromoplastic (TR6964|c11_g1_i1) were blasted against a database built based on marine microeukaryotic transcriptomes retrieved from the Marine

Microbial Eukaryote Transcriptome Sequencing Project (MOORE, Keeling et al., 2014) and therefore includes *L. danicus, L. aporus* and *L. hargravesii* transcriptomes; all of them were under the name of *L. danicus* though because they were submitted before the reappraisal of the genus by Nanjappa (2013). The database was built and kindly provided by Dr. Lisa Campbell (Texas A&M University, Department of Oceanography, personal communication). Nucleotide blast (blastn) was performed and the output was explored in MEGAN v.6.2.10 (Huson et al., 2007), using as taxonomic identifier the GI Mapping file also provided by Lisa Campbell. Based on these, 7 out of the 12 transcripts matched to *L. aporus* with a high score (HSFC1b, protein SYM1, HSF1, AOX4, HSFB-2a, TR6356|c2_g2_i1, TR7186|c6_g2_i10). No other good match to any other species was found.

The same transcripts related to heat and DNA integration were blasted against the transcriptomes of *L. danicus, L. hargravesii* and *L. convexus* that were produced in the purposes of the present thesis (see Chapter 4). None of them returned a good match either. Finally, a blast was also run against the metatranscriptomes from various time periods during the spring bloom of 2013 at LTER-MC station (Nanjappa et al., in preparation). Neither in this dataset was any good match returned for any of the transcripts implying a high specificity to the species or conditions of the query sequences.

3.4. Discussion

The objective of this part of the thesis was to identify expression differences for specific genes in response to temperature changes, which would ultimately help us understand the species adaptation to different environmental conditions and to mechanisms underlying diatom phenological patterns. To start with, the final transcriptome of *L. aporus* describes the functional capacity under specific thermal conditions since reads from samples grown under three temperature conditions were used for the assembly. The 64.5% of the *L. aporus* transcriptome assembly was annotated which is a percentage common for transcriptome annotation procedures. The *L. aporus* strains used for sequencing were selected in order to cover as high diversity as possible in terms of seasonality and physiological characteristics within the species.

The inference that the measured properties of a strain reflect those of the natural population from which it was isolated is a common assumption on which several phytoplankton studies rely on but it has been heavily questioned and criticized (Wood and Leatham, 1992; Lakeman et al., 2009). Therefore the effort of the present study was to acquire expression profiles at the different temperatures based on several strains that eventually would represent better the species' gene expression capability compared to gene expression data obtained from a single strain. Furthermore, *L. aporus* strains of all temperature groups have been chosen to be run on the same lane. As error profiles can be considerably different between sequencing runs and across subsections (e.g. Illumina lanes) within one sequencing run (Wolf, 2013), this was done to avoid that the differences detected in expression between treatment groups could be a result of possible differences in sequencing quality.

3.4.1. Differential expression analysis among temperatures

Based on the results of growth experiments described in Chapter 2, *L. aporus* was expected to have a stronger response at low rather than at high temperature. Indeed, the fact that there were more than 250 transcripts significantly differentially expressed when comparing high and low temperature, only nine transcripts between low and medium temperature and no significant differences between medium and high temperature means that *L. aporus* is mainly reacting to the low compared to the high temperature. The cold response of the species was more evident in the results of the k-means clustering. The expression profile did not differ dramatically between the high and medium temperature suggesting that 26 °C was a temperature that *L. aporus* could tolerate and under which the same functional level as at 19 °C was retained. All nine transcripts that were found significantly differentially expressed between low and medium temperature were also included in the significant ones between low and high temperature and followed the same regulation (up or down) of the low temperature condition. For the remaining 241 transcripts, their level of expression at high temperature was significantly different from that at low temperature but not at medium temperature. This agrees with the results of Chapter 2, which showed no significant differences of *L. aporus* growth between 19 °C and 26 °C.

Epigenetic changes are an important way in which genomes respond to environment, retaining the species plasticity and *L. aporus* followed this pattern as well. Reversible post-transcriptional modifications (DNA integration and dephosphorylation) were found significantly different between the different temperatures. The transcripts related to protein dephosphorylation and protein serine/threonine phosphatase activity were stress-activated kinases involved in the regulation of glucose and lipid metabolism as well as in many other cellular processes such as proliferation, division, survival and cell-cycle progression. Most of them were upregulated in low temperature condition. This is an indication of the stress that the species undergoes at low temperature when its metabolism and growth rate are really slow. The activation of transposable elements at low temperature could also be interpreted as a response to stress, as has been described in the introduction of the current chapter. In fact, manual annotation added even more transcripts to the TE related group reaching to sixteen in total. For four of them annotation was possible as Ty1-Copia type and two as Ty3/Gypsy type confirming the tendency of relatively high abundance of LTR retrotransposons in diatom genomes. Three other transcripts manually annotated were found to be probably homologous to the bacterial MAI genes related to biomineralization, which could have been acquired via horizontal gene transfer (Ullrich et al., 2005). It has been shown that MAI undergoes frequent rearrangements under physiological stress conditions including prolonged storage of bacterial cells at 4 °C or exposure to oxidative stress. In particular Ullrich et al. (2005) showed that MAI undergoes frequent transposition events. At the same time transposition of insertion elements can raise under stress conditions (Pfeifer and Blaseio, 1990; Mlouka et al., 2004). The overexpression of MAI at low temperature might mean that the mobilization of MAI in *L. aporus* contributes to genetic plasticity and finally adaptation to physiological stress in the same way transposable elements do. It must be noted that all the MAI related transcripts found significantly differentially expressed were actually highly expressed at low temperature in the recently isolated strains 1A1 and 3A6 while they were very low in all temperatures for the old B651 strain, which could be explained considering that long term cultivation may have caused the loss of this mechanism in the oldest strain. This expression

pattern was also the one mostly seen in the TE related transcripts and several of the rest of the stress-related transcripts mentioned above, an observation that led to the further investigation of the differences between strains which are discussed in the next section.

Chaperones, including HSPs and HSFs, could not be absent from the significant results of a temperature related gene expression study. As already mentioned in the introduction of this chapter, molecular chaperones are one of the most vital parts of the cell heat stress defense system and their role is to stabilize and fold proteins into their proper conformations (Richter et al, 2010; Feder and Hoffman, 1999). HSPs are produced in response to stressful conditions and it is now known that, in addition to heat, exposure to cold and UV light can lead to their expression as well (Matz et al., 1995; Cao et al., 1999). Another protein that could be essential during heat/ cold stress is SYM1, the stress-induced yeast ortholog of the mammalian Mpv17 mitochondrial inner membrane protein that is implicated in the metabolism of reactive oxygen species, ethanol metabolism and tolerance during heat shock (Trott and Morano, 2004). Furthermore, in a study on yeast Trott and Morano (2004) provided indications of a temperature-specific defect on SYM1 knockout cells that affects the activity of ALDH families. All the heat stress related transcripts mentioned so far, as well as ALDH, were significantly upregulated in *L. aporus* at low temperature. The only exceptions were heat shock factor protein 1 (HSF1) and heat stress transcription factor C-1b (HSFC1B) which were significantly upregulated in the high temperature. HSF1 is part of a complex with Hsp40/Hsp70 and Hsp90 and is inactive. Under heat shock, HSF1 is released from the complex by another complex including a translation elongation factor (eEF1A) which is also the key component regulating the actin cytoskeleton architecture in the cell. During the heat shock the general shutdown of protein synthesis leads to a collapse of the cytoskeleton which releases large amount of free eEF1A that now becomes available for interaction with HSF1 (Shamovsky and Nudler, 2008). HSF of the classes A and B are well established as regulators of thermal and non-thermal stress responses but the role of class C is unknown. The rice OsHsfC1b has been found to mediate salt stress tolerance, response to osmotic stress but also plant growth under non-stress conditions (Schmidt et al., 2012). HSFB2a has been found to elicit mild cell death

(moderate number of cells dying) in the leaves of the tobacco plant *Nicotiana benthamiana* (Zhu et al., 2012), whereas in *Arabidopsis thaliana* it is required for the development of female germline and the plant growth phenotype through temporal repression of vegetative growth during development (Wunderlich et al., 2014). The expression of HSFB2a is controlled by a heat-inducible long non-coding antisense RNA. Under heat stress the antisense regulation counteracts the effect of repression and restores growth and further development but at the same time leads to an impaired female gametophyte development. The heat induction of HSFB2a as well as of the antisense RNA depends on the presence of HSFA1a (Busch et al., 2005; Wunderlich et al., 2014). Under normal conditions, HSFA1a activity is repressed by negative regulatory mechanisms such as interaction with Hsp70 (Kim and Schöffl, 2002). In addition to heat stress, overexpression of HSFA1a has positive effects on stress tolerance to pH changes and to H_2O_2 in *Arabidopsis* (Liu et al., 2013; Qian et al., 2014). In *L. aporus,* HSFB2a and HSFA1a were among the genes significantly upregulated at low temperature so it is possible that also in diatoms HSFA1a expression is triggered under low temperature stress conditions leading to overexpression of the HSFB2a, implying a similar mechanism as in other organisms. Overall, although HSPs are now known to be expressed during many different types of stress beyond heat shock and even be involved in other non-stress induced functions, their significant upregulation is still a key part of temperature shock response induced primarily by HSFs and for that they should be considered an important element of the *L. aporus* response to low temperature.

Besides transposable elements, epigenetic changes and heat stress related proteins, an association between cold stress and oxidative stress has been already demonstrated in bacteria, cyanobacterial strains and fungi (Chattopadhyay et al., 2011; Smirnova et al., 2001; Liu et al., 2002; Hossain and Nakamoto, 2003; Gocheva et al., 2009). In those cases the expression of genes related to oxidative stress increased at low temperature, supporting the concept also mentioned in introduction of a temperature induced oxidative stress response, and the same pattern appeared in our study for *L. aporus*. Transcripts related to antioxidant-oxidoreductase activity were also significantly upregulated at low temperature. In particular, we could focus on LPOR, one

of the two unrelated Pchlide reductases (LPOR and DPOR) in the penultimate step of chlorophyll biosynthesis. It has been suggested that DPOR no longer operates in conditions where oxygenic photosynthesis is very active and cellular oxygen levels is high, while LPOR operates under both aerobic and anaerobic conditions, being able to compensate for DPOR loss (Yamazaki, 2006). The concentration of dissolved oxygen (DO) in water is influenced by temperature; the lower the temperature the higher the maximum DO concentration saturation. At 13 °C, an imbalance between reactive oxygen species and the ability of the biological system to detoxify the reactive intermediates or repair the resulting damage might occur in *L. aporus* and LPOR could be activated as a response to the disturbances in the normal redox state of the cell.

Two more evidences of the *L. aporus* stress reaction at 13 °C were protein degradation and the upregulation of AOX. Flick and Kaiser (2012) have illustrated concepts and mechanisms by which protein modification with ubiquitin and proteasomal degradation of key regulators ensures cellular integrity during stress situations. This was also one of the cases in *L. aporus* since protein degradation was the most enriched pathway in the differentially expressed genes between high and low temperature. AidB is one of the several genes involved in the SOS adaptive response (global response to DNA damage in which the cell cycle is arrested and DNA repair and mutagenesis are induced) to DNA alkylation damage, the expression of which is activated by Ada protein. It has been proposed that aidB directly destroys DNA alkylating agents such as nitrosoguanidines (nitrosated amides) or their reaction intermediates (Annocript annotation, see Materials and Methods). The transcripts of this pathway were all significantly upregulated at low temperature implying changes in metabolism in order to cope with the unfavorable environmental conditions. AOX is one of the two terminal oxidases (the alternative one, as its name implies) of the mitochondrial electron transport chain and it can dramatically reduce the energy (ATP) yield of respiration compared to the cyt oxidase. The expression of AOX is induced when plants are exposed to a variety of stresses including oxidative stress, chilling pathogen attack senescence and in rice in particular the transcript levels of the alternative oxidase are increased by low temperature. So, it is an important mitochondrial component of the plant stress

response while it also holds the ability of maintaining metabolic and signaling homeostasis; a link between mitochondrial function, signal transduction and acclimatization to stress (Vanlerberghe, 2013). AOX, but also Aldh6a1, RaiA, FDHs and YdcS which are all related to stress conditions (described also in section 3.3.1), were upregulated at low temperature in *L. aporus*.

3.4.2. Differential expression analysis among strains

The hypothesis that stimulated the comparison among strains was the one already formulated in Chapter 2 about the different behavior of the older strain B651 possibly due to in-culture evolution, which was also evident in the results of the differential expression analysis among temperatures. Therefore, the differential expression analysis among strains was designed and it did produce some interesting results. First of all, the number of the significantly differentially expressed genes detected in this analysis was much higher than in the temperature DE analysis and in particular there were many more genes found significantly different between B651 and the other two strains than between 1A1 and 3A6. This is a first confirmation of the different behavior of strain B651 compared to the other two in terms of expression. Then, the TEs' informative role was confirmed by the high enrichment of DNA integration terms in all three strain pairs; yet, the highest percentage was present in the B651 - 1A1 pair and the lowest one in the 1A1 – 3A6. Thus, an extra clue was added regarding the involvement of TEs in *L. aporus* evolution. TEs can be activated due to their ability to mask themselves and parasitize the genome taking advantage of the cell stress responses but ultimately they are kept active because of the benefits they offer to the host. As Cavrak et al. (2014) conclude in their review "evolution has the last word". Therefore, TEs might serve as important tools for the adaptation and evolution of species and we consider them to be a significant point to focus on in our own analysis on Leptocylindraceae reaction to the different environments.

The involvement of TEs as well as other stress related genes in the B651 differential expression compared to the other two strains does indicate a different response to low temperature but it is impossible to say with certainty if it is due to the possible in-culture evolution, the high functional intraspecific variability of *L. aporus* species or the fact that is the only representative of a putative

summer population. It has been argued that phytoplankton clonal isolates show evolutionary changes as they accumulate mutations and adapt to culture conditions (Lakeman et al., 2009). Populations initiated as clonal isolates have also been shown to develop cryptic genetic variation in sensitivity to heavy metals (Lakeman and Cattolico, 2007), competitive ability (Costas et al., 1998) and growth rate and grazer defenses (Yoshida et al., 2004; Becks et al., 2010). Strains of the green alga *Chlorella* showed significant interclonal variation in defenses against rotifer predation, traded-off with competitive ability and growth rate, after more than 50 years of maintenance in culture collection in the complete absence of grazer (Yoshida et al., 2004; Meyer et al., 2006). Demott and McKinney (2015) showed a decline in digestion defenses and an increase of growth rate of a strain of the green alga *Oocystis* after 3 years of culture. The loss of grazer defenses was suggested to be a gradual process, dependent on the number of cell divisions over several years and consistent with a quantitative genetic trait determined by mutations at many gene loci, each with a small, accumulative effect (Houle, 1992; Lakeman et al., 2009; Demott and McKinney, 2015). Such evolutionary change and genetic diversity complicate the interpretation of laboratory experiments. The rate of evolution in clonal cultures presumably depends on culture conditions, population size, generation time and the traits being measured and therefore appears to be highly variable (Lakeman et al., 2009). In a clonal, asexual population, such as in *L. aporus,* initial slow changes are predicted by theoretical modeling as mutations gradually accumulate, followed by more rapid evolution within a few hundred days (Lynch et al., 1991). Although large population size and short generation times in algal cultures favor evolutionary change, the timing of specific phenotypic changes is hard to predict. A quite notable example is that of a dinoflagellate subculture that stopped producing saxitoxin around 40 years after isolation while another subculture kept under similar conditions continued to produce the toxin (Martins et al., 2004). At the other extreme, clonal isolates of a dinoflagellate evolved improved interclonal competitive ability during experiments lasting only 5 weeks (Costas et al., 1998). Based on the above, the effect of in-culture evolution cannot be ignored but it is hard to predict to what extent it affects each strain separately. In the case of B651, the different behavior is assumed to be mainly a result

of the in-culture evolution rather than of its summer origin. This hypothesis is also supported by Chapter 2 results that showed two additional 'old' strains to be different regarding the extra time needed to acclimatize, although these were isolated in autumn.

3.4.3. Transposable elements in *L. aporus*

To start with, all TE related transposons that were significantly differentially expressed in the temperature based DE analysis were also significantly differentially expressed in the strain based DE analysis. This is a first sign that the TE related transcripts both contribute to the variability among strains and are involved in adaptation to changing thermal environments. Indeed the expression pattern of the majority of the significantly differentially expressed TE-related transcripts is differentiated based on the stress factor, which is low temperature in our case, but could also be linked to adaptation which is reflected by the different behavior of the strain B651, maintained for years in culture. The expression of all the selected TE transcripts was very low or even absent in B651 except for one (TR6586), which followed the opposite trend, being almost absent in 1A1 and 3A6 but considerably expressed in B651. Focusing on the B651 case, the TE transcripts could have been active when the strain was experiencing the variability of the natural environment but eventually silenced or constrained by the genome after the cells were isolated and kept in stable environmental conditions for years. The TE that showed an opposite behavior might be an interesting example of a possible inverse procedure where a TE is activated/ enhanced in order to maintain a stress response despite the stabilization of the environment and the absence of stress. A hypothesis is that cells silence the majority of the TEs which are not useful anymore due to the absence of environmental variability but they still keep as a backup few TEs, the expression of which is enhanced instead. This trade-off should be energetically beneficial for the cell to select it for. In addition, the co-expression of HSFA1a and transposon related genes might be evidence of a mechanism similar to the one found for the *ONSEN* transposable element in *Arabidopsis,* where HSFA1a was proved to be necessary for the heat-induced expression of the TE (Cavrak et al., 2014).

The qRT-PCR results of the validation set basically confirmed the RNA sequencing results. Regarding B651 lower expression tendency, the pattern is confirmed for TR6586 and possibly for TR6506_i3 and TR7186. On the other hand, when the RNA-seq results are compared with qRT-PCR results of the exploration set produced from the 2016 samples, there seems to be a big discrepancy, in many cases even to the opposite direction. The only gene also confirmed in the 2016 samples is SYM1. The transcripts were in fact not at all or expressed in a very low level in the sample that was acclimatized for the shorter time (1189B3). Considering these results altogether, there are two possible explanations:

- Transcripts are in reality reacting randomly to the different conditions to which the cultures were subjected and there is no actual connection with temperature, stress or adaptation (except for SYM1 and temperature).
- In the two different periods when the 2015 (validation set) and 2016 (exploration set) experiments were conducted, the physiology and therefore the condition and reactions of the strains were different. This is due to some factor other than the stress condition per se.

In support to the latter hypothesis, the growth rates during the two periods were indeed different. Strains during 2015 seemed to show a lower growth rate than during 2016, when prepared for qRT-PCR. The main difference between the two periods was the duration of acclimatization. If this was the case, then 2015 samples could be assumed to be more stressed due to the longer acclimatization period. The stress during 2015 could be possibly higher also due to a contamination by a heterotrophic flagellate that almost all samples underwent during this period. Before accepting this scenario, all other possibilities should be examined. The other scenario was that the strains' different behavior was not really an outcome of the acclimatization duration but of the strain specific reaction due to their seasonal id (isolation date) or the season during which the experiment was performed. If there is any effect of the seasons on diatoms in the same way there is in plants (endogenous clock, also mentioned in Introduction) the isolation or filtration dates should be correlated to the growth rates. However, the statistical tests' results

did not show any real correlation between the growth rate and the season of isolation of the strains. In the filtration dates though, a seasonal pattern could be seen; March was found different compared to January and February.

Overall it seems that the difference of the growth rates could be both related to the acclimatization period but also related to the season. Nevertheless, the gene expression levels point towards acclimatization to have the strongest effect on the strain expression response. All the transposon related transcripts were confirmed regarding their response to the cold stress. The pattern for TR_6586, and to some extent for TR6506_i3 and TR_7186, was also confirmed regarding the different behavior of the older strain. Although this result could be an evidence for the presence in *L. aporus* of specific transposable elements that are involved in phenotypic plasticity of the strains, the contrasting expression patterns noticed in the samples that underwent a shorter period of acclimatization indicate an essential role of the exposure time to the stressing factor. Indeed, the different acclimatization to a stressful environment can lead to distinct reactions in algae, causing either the reduction of the growth rate in the long term acclimatization compared to the short one due to energy allocation from growth extension to maintaining the cell structural integrity, or a restoration of the growth rate after a longer exposure to the stress due to adaptive plasticity (Ragazzola et al., 2013; Schlüter et al., 2014). Longer term experiments are suggested to provide a better understanding of the species response to environmental changes that tend to persist such as global warming or ocean acidification (Form and Riebesell, 2012). As described in section 1.5, phenotypic variability within populations, a result of high level of inherited phenotypic plasticity, plays an important role in the adaptation of the plankton to changing environments, potentially constraining short-term effects and forming the bases for selection. In the case of *L. aporus*, the long-term exposure to the stress factor seems to affect negatively the growth rate but activates most of the stress-related transposons, an energy allocation perhaps similar to that suggested for the algae *Lithothamnion glaciale* (Ragazzola et al., 2013). On the other hand, differences in transposon expression levels were also recorded among the newly isolated strains, where the acclimatization of 1188A1 was

only 10-20 days longer compared to the other two. However, there are strains in the validation set that differed remarkably regarding their acclimatization time (around 100 days) without any clear influence in gene expression results. Therefore it seems that for these transposons there is a temporal threshold of about 10-20 days, after which they are activated. After this threshold is passed the expression changes are not so profound and the cell state is stabilized.

On the other hand, significant relationships revealed between strain-specific growth rates and the impact of stress on growth show that the response to the short and long-term changing environment is not only species but also strain specific (Kremp et al., 2012). In the current experiment, the strain specific reaction was more complex and harder to interpret since the strains isolated more recently were also the ones that underwent the shorter acclimatization. The strain effect would be clearer if all the strains had been acclimatized for the same period of time. In any case, each response pattern should be always treated as strain specific and the general conclusion for the species should be kept in a flexible frame.

The temperature related transcripts did not show any of the sample dependent patterns that were observed for the TE transcripts. The highly conserved domains in the HSF family might have affected the final expression detected by the qPCR for the specific two HSFs tested here. SYM1 was the only one that was confirmed in all the low temperature samples of all acclimatization groups. This confirms that SYM1 in *L. aporus* is implicated in the cold and oxidative stress response since this protein, as already mentioned, is involved in the metabolism of reactive oxygen species and tolerance during heat shock. So far, little mechanistic information exist for this protein but the hypothesis for its function in yeast is that it holds a pore or channel-like activity in the mitochondrial inner membrane required for the transport of small molecules into or out of the mitochondrial matrix under heat shock conditions (Trott and Moranno, 2004). The function in diatoms could be similar or slightly different. Yet, the high stability of this protein expression through all strains and conditions makes it a good candidate marker for thermal stress in Leptocylindraceae and possible in other diatom species with similar physiology.

SYM1 was blasted against NCBI and the Mpv17 domain was found conserved, though with a low similarity (ca 40%), in the diatoms *Thalassiosira oceanica* and *Thalassiosira pseudonana* and the coccolithophore *Emiliana huxleyi.*

The specific genes implicated in the response of *L. aporus* to cold could be genes that are conserved and utilized by other species for similar stress responses or genes characteristic of *L. aporus.* The investigation of this scenario could provide some answers on one of the big questions of the thesis regarding the reaction of co-occurring species under the same environment. The results of the blast search of all the genes of interest against the MOORE foundation database, the individual transcriptomes of other *Leptocylindrus* species produced in the frame of this thesis and the metatranscriptome from the spring bloom of 2013 in MareChiara station proved that the transcripts of interest investigated here are unique to *L. aporus* so far. They could be either species-specific or only strictly activated by the temperature change to very low or high values. The *Leptocylindrus* transcriptomes of the other species were all acquired at 19 – 20 °C and the environmental sample was taken during the course of a late spring bloom where the transcripts might not reach such high levels of expression due to the absence of the appropriate trigger (low temperature). Considering that none of the databases against which they were blasted represent actual stressful conditions, this hypothesis is possible and so the genes might be there but not detected.

3.4.4. Conclusion

The transcriptome analysis of three *L. aporus* strains under three different temperatures revealed that the species is stressed at 13 °C. Cold stress conditions in *L. aporus* lead to the activation of pathways related to temperature and oxidative stress but also to the activation of many transposable elements. However, the differences among strains were as important, if not of greater importance, compared to the differences among temperature conditions. In particular, one of the strains, the oldest one, showed signs of diversification possibly related to in-culture evolution. In the between-strains transcriptomic analysis, transposable elements held again a

central role. All of them were of the LTR retrotransposon superfamily, as expected from previous studies on diatoms' transposable elements. At this point the central hypothesis was formulated:

- The TE related transposons are activated after cold stress but they are silenced when the cells remain for a long period in the same environmental conditions. This would mean that TEs in *L. aporus* provide the phenotypic plasticity required in a changing environment that can lead to genetic diversity within a species. Selection acts on the diversity created by TEs and adapted populations are established and this would be the case of the older strains.

At least three of the transposable elements investigated could be considered as important candidates for their implication in the cold stress response and possibly in the species adaptive evolution as well. However, one of them, TR6586, is a transposon that shows a particular expression pattern since it is highly expressed in the older strain in contrast to the rest TEs; possibly it is a selected TE for enhanced activity in case of an unexpected cold stress which might occur despite the phenomenal stabilization of the environment. Nevertheless, the activity of all TEs was found to be dependent on the persistence of the stress factor and they were activated after at least 10 to 20 days of exposure to low temperature. It is accepted based on the results of the growth experiments (Chapter 2) that the strains of a species can be impacted in very different ways by stressors but it was not possible to identify clear strain specific responses, if any, in the current transcriptomic experiment. The significant upregulation of transposable elements in *L. aporus* could be an important finding since it might add one more diatom species to the ones already studied regarding the role of TEs in their adaptation to changing environments. The current results support the theory of the high importance of transposons in diatoms which states that, considering TEs are actually a mechanism of phenotypic plasticity, their activation after a certain time of highly changing environmental conditions can induce the genetic diversity that allows diatoms to adapt so successfully to so many environments.

Furthermore, SYM1 was verified as a cold response protein which also contributes to the thermal shock tolerance and it seems to be necessary already at the 10 days acclimatization. Its high

positive consistency in the results could lead towards the further investigation of its exact functions in diatoms and finally make it a reliable reference gene for cold stress identification.

Chapter 4. Comparative Transcriptomics in *Leptocylindrus* species

4.1. Introduction

Diatoms are among the most diverse groups of phytoplankton in the ocean but, despite their widely recognized importance, little is known about the extent of their functional diversity, how this diversity shapes and reflects their different distribution patterns and the mechanisms through which it is maintained or increased over time. The characterization of the functional diversity of diatoms can offer essential information and help to determine whether or not the marked diversity in gene-specific responses translates directly to significantly different ecological consequences in the field. To that end, analyzing the functional characteristics of different species through the comparison of the quality and the quantity of mRNA they produce, namely their gene expression profile, can help identify how these species adapt to the environment where they live.

Before the development of HTS technologies, other techniques such as expressed sequence tags (ESTs) and microarrays were used in most gene expression studies. These approaches have now been replaced by deep sequencing, due to the superiority of the latter. The power of sequencing RNA is the combination of the twin aspects of discovery and quantification in a single high-throughput sequencing called RNA sequencing (RNA-seq). Whole transcriptomes acquired by RNA-seq are often compared between or among different species and this is called comparative transcriptomics. Using comparative transcriptomic analysis, we might be able to obtain a preliminary insight on molecular toolkits specific to each species and even identify physiological and metabolic differences amongst them.

In order for the interspecific comparison to be possible, the RNA-seq data should be indeed comparable among species; expression can be statistically compared only for genes that are of common ancestry and retain the same function (orthologs). The most important issue is that a one-to-one correspondence between genes from different species does not exist due to evolutionary events such as gene duplication and recombination which create complex relations between genes (Kristensen et al., 2011). So even for orthologs (genes from different species with a shared common ancestry) such a correspondence is not obvious since in-paralogs (one or more

additional gene copies resulted by gene duplication after speciation) are a possible case. The function of the paralogs tends to diverge over time and have in general a high gene expression diversity compared to the single-copy genes (Gu et al., 2004; Studer and Robinson-Rechavi, 2009; Kristiansson et al., 2013).

In diatoms, similar whole transcriptome functional comparisons have indicated a comparable set of functions between the species with small differences in specific pathways (Di Dato et al., 2015; Bender et al., 2014). Despite the overall similarity, it was also possible to identify species-specific transcripts. The characterization of both the common and distinct responses for each species can help to understand and better predict which diatoms bloom under which sets of environmental factors. Each species might maintain different physiologies from one another in how each undergoes -or not- sexual reproduction, responds to micronutrient availability and nutrient storage in ocean environments etc. (Strzepek and Harrison, 2004; Peers and Price, 2006; Sims et al., 2006; Armbrust, 2009). Bender et al. (2014) examined the transcriptional response of three diatoms – *Thalassiosira pseudonana, Fragilariopsis cylindrus and Pseudo-nitzschia multiseries* – to the onset of nitrate limitation of growth in order to investigate the ramifications of between-species diversity. They concluded that the observed diversity was a result of four general mechanisms:

1. Genes specific to each diatom species. However, distantly related genes may encode diverged proteins that carry out similar metabolic functions in diatoms. This means that different transcriptional responses could nevertheless lead to similar functional reactions (Thompson et al., 2011).
2. Multi-copy gene families to which a small percentage of proteins within each diatom belong to. These families commonly display different transcriptional patterns among different diatoms but also within individual species.
3. Differential regulation of separate components of the same metabolic pathway which may be a result of post-transcriptional and post-translational modification for selected genes or proteins respectively.

4. Variations in the regulatory networks that differentially control gene expression among the different species.

In challenging environments the species-specific physiological capabilities and adaptations are highlighted (Thompson et al., 2011). So when different species or strains are subjected to similar perturbations, e.g. extreme temperature, and a comparison in responses follows, a greater molecular diversity and larger differences in the gene expression levels is expected to be detected than when all individuals are kept in tolerable conditions. However, the examination of gene expression under non – stressful conditions should not be neglected since it could also provide valuable insight into the functional diversity of species under normal conditions.

The considerable amount of sequence information that is produced by RNA-seq can be combined with phylogenetic principles in an attempt to make further sense of the data. This kind of approach is called phylogenomics and is an intersection of the fields of evolution/ phylogenetics and genomics (Eisen and Fraser, 2003). It has been already largely applied in diatoms leading to significant conclusions, such as the red and green algal origin of the diatom membrane transporters (Chan et al., 2011; Deschamps and Moreira, 2012; Derelle et al., 2016). Phylogenetics is based on the identification of homologous characters (morphological structures, ultrastructural characteristics of cells, biochemical pathways, genes, amino acids or nucleotides) shared among different organisms so through their comparison and the use of reconstruction methods evolutionary relationships and phylogenetic trees are inferred. Whereas some few genes show a high degree of conservation across all organisms e.g. small subunit ribosomal RNA (SSU rRNA) and are hence preferable, many other genes show topological conflicts among phylogenies. Each molecular phylogenetic tree is based on differences among molecules and it is not necessarily a species tree. Aminoacid sequences of regulatory proteins or enzymes commonly diverge to a high extent, except for residues absolutely required for activity, while rRNA genes are the most conservative large sequences in nature and therefore fit the criteria for representing the actual species phylogeny better than any other genes (Woese, 1987). In that sense, phylogenetic trees made based on other genes of the central nucleic acid-based information transfer process, such

as RNA or DNA polymerase, are consistent with the rRNA trees while metabolic processes that respond to the environment may or may not track the rRNA (Brown et al., 2001; Pace, 2009). Therefore the possibility of combining many genes together could solve problems such as lateral gene transfer, convergence and varying rates of evolution for different genes. Using entire genomes or transcriptomes should bypass these anomalies since the pattern of evolution is indicated by the majority of the data (Delsuc et al., 2005; Jeffroy et al., 2006). Summing up, phylogenomics is a new field that has arose along with the fields of genomics and transcriptomics and it could be highly promising when the selection of the genes is restricted to the ones that contain minimal non-phylogenetic signals in order to take full advantage of the method and reduce incongruence.

Within the phylogenomics field, an increasingly popular analysis that can be applied on the results of HTS experiments is the variant calling. A range of computational methods is utilized in order to identify the existence of Single Nucleotide Polymorphism (SNP) and small insertions and deletions (indels) in the studied sample compared to the assembled transcriptome. The next steps of variant calling are (i) the stratification of the variants based on their impact on the protein sequence such as a synonymous SNP when located within the coding region and a non-synonymous SNP otherwise, or a frameshift mutation that led to the gain of a stop codon, and the (ii) annotation of the transcripts significantly affected by the variants. In that way, specific genetic polymorphisms can be associated with a species and further investigated regarding their origin and evolution, with no need of sequencing the entire or part of the genome (Lopez-Maestre et al., 2016). Apparently, in the absence of a reference genome only the polymorphisms of transcribed regions can be targeted but these regions arguably correspond to those with a more direct functional impact since RNA-seq data mirror gene expression, the most basic molecular phenotype.

The *Leptocylindrus* species used in the current analysis offer a fine opportunity to link different phenotypes with RNA-seq and phylogenomics results since *L. danicus* and *L. hargravesii* are genetically closer than the other species and they are undistinguishable under the light

microscopy while *L. aporus* and *L. convexus* are also genetically closer but their discrimination is much easier under the microscope; additionally, *L. danicus* and *L. hargravesii,* but not the other species, are known to undergo sexual reproduction with flagellated gametes and spore formation from auxospores. All species show distinct seasonality with some overlapping periods: *L. danicus* is an almost year-round species with low numbers in mid-July to August; *L. aporus* is detected as abundant when *L. danicus* is not but with very low density in the winter months; *L. convexus* is a half-year species, with low numbers in January-February and higher in spring months up until July; *L. hargravesii* is present mainly in summer and beginning of autumn but also in January – April in lower numbers. It will be interesting to see how this complexity of molecular, morphological and seasonal characteristics is depicted in the functional patterns of each species and identify different and/ or evolutionary conserved transcriptional responses. In particular:

- We aimed at understanding how genetically variable each species is and around which functions this variability is centered. Therefore, the genetic variability among strains and species was explored through a variant calling analysis. The exact effect of the variants on the protein sequence was also determined and finally the specific functions mostly affected by the polymorphism were detected for each species through a GO enrichment analysis.
- A differential expression analysis among strains for each species was aimed at exploring the diversity in gene expression and the related functions within the species.
- The identification of orthologous genes among all species and the following differential expression analysis was aimed at exploring the diversity in gene expression and the related functions among the species.
- Finally, due to the enormously large data produced it was impossible to investigate all differentially expressed genes or manually annotate transcripts that missed annotation. Instead the investigation of expression was specified and focused on the concepts related to the general aim of the thesis and the genes that turned out to be of interest in the transcriptomic analysis of *L. aporus* in the previous chapter.

4.2. Materials and Methods

Based on the results of the molecular characterization of a high number of strains (see Chapter 2), strains of each species were selected so that, in combination with strains already available in the SZN collection, they would cover as high diversity as possible in terms of seasonality and possible physiological characteristics (Table 4.2.1). Based on the ITS marker, a maximum likelihood tree with bootstrap and Kimura-2 parameter model was built and intra-species molecular similarity was also taken into consideration.

Table 4.2.1 Strains selected from each Leptocylindraceae species for RNA sequencing and corresponding dates of filtration for RNA extraction.

#	Isolation Date	Strain Code	Species	Filtration Date
1	20/12/2013	1A1	*L. aporus*	16/12/2014
2	28/01/2014	3A6	*L. aporus*	17/12/2014
3	21/8/2010	B651	*L. aporus*	01/02/2015
4	14/01/2014	1089-07	*L. convexus*	06/04/2015
5	21/12/2010	B768	*L. convexus*	16/04/2015
6	23/09/2014	1123B2	*L. convexus*	06/04/2015
7	14/01/2014	1089-21	*L. hargravesii*	05/03/2015
8	05/02/2014	3B6	*L. hargravesii*	15/03/2015
9	19/02/2014	4D4	*L. hargravesii*	13/03/2015
10	13/02/2014	4B6	*L. danicus*	19/05/2015
11	14/01/2014	1089-17	*L. danicus*	31/01/2015
12	15/06/2010	B650	*L. danicus*	14/05/2015

The selected strains were grown under the same conditions at 19°C, a light intensity of 100 µmol photons m^{-2} sec^{-1} and a photoperiod of L:D, 12:12. The environmental conditions were stable while cultures were kept at exponential phase in 100 ml of K + Si medium and finally transferred to 1 liter (1L) when reached a concentration of 2,000 cells/ml. RNA extraction and sequencing were performed following the same protocols and by the same services described in Chapter 3. All samples except *L. aporus* were placed on the same lane for sequencing. Samples run in the same lane avoid that the differences detected in expression between treatment groups could be a result of possible differences in sequencing quality.

After the acquirement of the sequences, the data were sent to Sequentia Biotech (Barcelona) for the downstream analysis. An initial quality check was performed on the raw sequencing data, removing low quality portions while preserving the longest high quality part of a HTS read. The

minimum length established was 35 bp and the quality score 25, which increases the quality and reliability of the analysis. *FastQC* analyses were performed before and after the trimming of the reads (removal of adapters) in order to have a quality control of the high throughput sequence data. The steps of the main analysis follow in details:

• **De novo transcriptome assembly for each species**. At this step a transcriptome for each species was assembled without the aid of a reference genome. Assembling the transcriptome of each *Leptocylindrus* species was the first step in developing large scale genetic information that allowed us to further study the recurrent phenotypic evolution across the samples. The high quality reads obtained after trimming were post-processed more in-depth. First, all reads of the samples belonging to the same species were joined in order to increase the transcriptome assembly accuracy and coverage. Then, an *in silico* read normalization was carried out. Read normalization reduces the redundancy of the dataset thus speeding the analysis and increasing the quality of the assembly. Finally, the obtained normalized datasets were aligned against *Homo sapiens* reference genome (GRCh38) in order to remove contaminants. All the reads mapping against GRCh38 genome, were discarded from downstream analysis. At the end of the read processing, the total amount of reads obtained for *L. aporus, L. convexus, L. danicus* and *L. hargravesii* were used as input to perform transcriptome assembly. Different analyses and comparisons were conducted using two of the most outstanding tools for de novo assembly, Trinity (version 2.1.1) and Trans-ABySS 1.5.3 (Simpson et al., 2009). The best assembly was obtained by merging the assemblies produced by both assemblers separately. Only those transcripts bigger than 200bp length were assembled; smaller ones were considered as contaminations, or assembly artefacts. Then, the transcript redundancy was removed with CD-HIT-EST obtaining a "raw" assembly. CD-HIT-EST is a widely used program for clustering similar proteins into cluster that meet a similarity threshold.

• **Comprehensive quality analysis and filtering of the assembled transcript sequences.** In order to have a general overview about the quality of the assembly, a tool called Transrate v1.0.0 (Smith-Unna, preprint) was used. Transrate is a software for de novo transcriptome assembly

quality analysis. It examines the assembly in detail, reporting quality scores for assembled sequences. Furthermore it filters out the bad sequences from the assembly, i.e. those sequences that might be assembly artifacts. All the transcriptomes were analysed using Transrate, except the one from *L. aporus*. The *L. aporus* RNA-Seq reads were shorter than 40bp, and Transrate needs larger reads as input. Therefore, *L. aporus* assembly was checked using a tool called RSEM-EVAL v.1.9 (Li and Dewey, 2011), which is another de novo transcriptome assembly evaluator. HGS data could suffer from contamination of organisms that are not the actual target of the experiment (apart from human), so the sequences of the transcripts were blasted against the NCBI database in order to remove possible contaminations (all no plant or diatomea hits).

• **Functional annotation of high quality transcriptome**. The high quality transcripts were translated into proteins with the TransDecoder tool (http://transdecoder.github.io/). The sequences of the assembled transcripts were screened for open reading frames (ORFs) in order to predict the amino acid sequence of proteins derived from them. When multiple translations were possible, the priority was set in order to get the longest complete ORF, when a complete ORF was not detected the longest sequence was kept. The coding sequences obtained from TransDecoder were functionally annotated with the InterPro database. InterProScan is the software package that allows sequences to be scanned against InterPro signatures. InterPro provides functional analysis of proteins and classifies them into families. This step also let the production of Gene Ontology (GO) and KEGG annotations. In addition to InterPro annotations, using protein sequences, a blastp analysis was performed against NCBI in order to functionally annotate the proteins by sequence similarity. The gene names of the most similar NCBI protein hit were retrieved (all of them having an evalue ≤ 0.01).

• **Identification of variants across different strains within each species (Single Nucleotide Polymorphism (SNP) and small insertions and deletions (indels)) and genomic variant filtering.** Following the assembly and annotation of the transcriptomes, variants across the different strains of species were identified and annotated. The variant calling pipeline involves the identification of

variants present in the studied sample in comparison to the assembled transcriptome. This process is divided into further steps:

1. Mapping and data pre-processing
2. Variant Calling
3. Variant Filtering

The first step includes the alignment of the trimmed reads against the assembled transcriptome using the latest version of Spliced Transcripts Alignment to Reference (STAR) software, v2.5.1 (Dobin et al., 2012). STAR outperforms other aligners by more than a factor of 50 in mapping speed of RNA-Seq reads, while at the same time improving alignment sensitivity and precision. To check the mapping quality of the alignment, SAMStat software (Lassman et al., 2011) was used. The resulting alignments obtained after mapping were pre-processed to make them adequate for variant calling analysis: using SAMtools 1.2 the reads mapping with a mapping quality less than 30 and not properly mapped were removed from downstream analysis. Once the data were pre-processed, variant discovery process was carried out, i.e. the sites where the data displays variation relative to the reference sequence were identified, and genotypes for each sample at that site were calculated. For this scope, a pipeline called SUPER v4.0 was used (https://sourceforge.net/projects/superw/). Simply Unified Pair-End Read (SUPER) workflow is a dynamic and fast tool to identify sequence variation such as SNPs, DIPs and Structural variations (SVs) developed by Sequentia Biotech team. Subsequently, several filters were applied in order to reduce the amount of false positives and obtain the most accurate and reliable variants:

- Variant Quality: Those variants with a quality less than 30 were removed from downstream analysis.
- Genotype Depth: SNP/indels with less than 6 reads of coverage were removed from downstream analysis.

• **Accurate variant annotation and effect prediction**. As soon as the variants were established, their annotation followed. There are many different types of information that could be associated with variants. First, we focused on the most fundamental level of variant annotation, which is

categorizing each variant based on its relationship to coding sequences in the transcriptome and how it may change the coding sequence and affect the gene product. SnpEff tool was chosen as the best software to perform this analysis (Cingolani et al., 2012). SnpEff v4.1b is a genetic variant annotation and effect prediction toolbox. It annotates and predicts the effects of variants on genes (such as amino acid changes). It classifies the variants as intergenic, intronic, nonsynonymous SNP, frameshift deletion, large-scale duplication, etc. (Sequence Ontology terms, http://www.sequenceontology.org/), and based on this annotation, an assessment of the putative impact of the variant follows (Table 4.2.2). A high impact variant is assumed to have high (disruptive) impact in the protein, probably causing protein truncation, loss of function or triggering nonsense mediated decay; a moderate impact variant is a non-disruptive variant that might change protein effectiveness, a low impact variant is assumed to be mostly harmless or unlikely to change protein behavior and a modifier is usually a non-coding variant or a variant affecting non-coding genes, where predictions are difficult or there is no evidence of impact

Table 4.2.2. Putative impact for Sequence Ontology terms often used in functional annotations.

Putative Impact	Sequence Ontology term
HIGH	chromosome number variation
HIGH	exon loss variant
HIGH	frameshift variant
HIGH	rare amino acid variant
HIGH	splice acceptor variant
HIGH	splice donor variant
HIGH	start lost
HIGH	stop gained
HIGH	stop lost
HIGH	transcript ablation
MODERATE	3 prime UTR truncation + exon loss
MODERATE	5 prime UTR truncation + exon loss variant
MODERATE	coding sequence variant
MODERATE	disruptive inframe deletion
MODERATE	disruptive inframe insertion
MODERATE	inframe deletion
MODERATE	inframe insertion
MODERATE	missense variant
MODERATE	regulatory region ablation
MODERATE	splice region variant
MODERATE	TFBS ablation
LOW	5 prime UTR premature start codon gain variant
LOW	initiator codon variant
LOW	splice region variant
LOW	start retained
LOW	stop retained variant

LOW	synonymous variant
MODIFIER	3 prime UTR variant
MODIFIER	5 prime UTR variant
MODIFIER	coding sequence variant
MODIFIER	conserved intergenic variant
MODIFIER	conserved intron variant
MODIFIER	downstream gene variant
MODIFIER	exon variant
MODIFIER	feature elongation
MODIFIER	feature truncation
MODIFIER	gene variant
MODIFIER	intergenic region
MODIFIER	intragenic variant
MODIFIER	intron variant
MODIFIER	mature miRNA variant
MODIFIER	miRNA
MODIFIER	NMD transcript variant
MODIFIER	non coding transcript exon variant
MODIFIER	non coding transcript variant
MODIFIER	regulatory region amplification
MODIFIER	regulatory region variant
MODIFIER	TF binding site variant
MODIFIER	TFBS amplification
MODIFIER	transcript amplification
MODIFIER	transcript variant
MODIFIER	upstream gene variant

As a last step for the variant analysis, a GO enrichment analysis was carried out for those transcript significantly affected by high impact variants. For this purpose, agriGO tool was used (Du et al., 2010), through which the identification of significantly enriched GO terms among the selected set of transcripts with accurate statistical methods was possible.

• **Differential expression analysis across species and among individuals of the same species**. Differential expression analysis has been performed with NOIseq R packages, a nonparametric approach that does not require any preset distributional assumptions as it creates an empirical distribution from the available data. NOIseq allows DE analysis without biological replicates. Prior to the analysis, the “Trimmed Means of M-values” (TMM) normalization strategy was used. A first filtering procedure has been performed in *L. convexus* and *L. hargravesii* analysis: the transcripts with less than one read count in only one sample were removed. EdgeR package was used for this scope. Differentially expressed transcripts were discovered by pairwise comparison of the three stains in each species. Transcripts are considered significantly differentially expressed if the false discovery rate (FDR) of the statistical test is less than 0.05. Finally, Gene Ontology Enrichment

Analysis (GOEA) was carried out. Enriched GO terms were identified among the differential expressed set of loci with accurate statistical methods. This analysis was conducted on the differential expressed transcripts identified on each comparison of the previous step (DE-analysis) and was carried out on the lists of up and down regulated transcripts separately.

• **Identification of orthologous transcripts across the species**. As a final analysis, orthologous transcripts were identified across the four species and an interspecies differential expression analysis along with a corresponding GOEA was carried out. The strategy chosen for the identification of orthologous sequences is based on the Reciprocal BLAST Hits (RBH). Essentially, a RBH is found when the proteins encoded by two genes (or transcripts), each in a different genome, find each other as the best scoring match in the other genome. This method was originally developed to identify orthologous sequences between two species. In this case, the RBH method was adapted to identify conserved sequences not only among pairs of species but also across all the species or across groups of three. Once the groups of orthologous transcripts were defined, all the sequences of the orthologous genes across the species were used for a neighbor joining tree. The method was based on an alignment free calculation followed by phylogenetic reconstruction with PHYML (Guindon et al., 2010). A differential expression analysis across the different species was also performed. A first attempt was made to map the reads of each species using one reference sequence. However, the number of mapping reads was too low. For this reason the expression values calculated for each species using its own reference were used for the analysis. The differential expression analysis was performed with NOIseq R package. Using the transcripts conserved across the four species, the following comparisons were made:

- *L. aporus* vs *L. convexus*
- *L. aporus* vs *L. danicus*
- *L. aporus* vs *L. hargravesii*
- *L. convexus* vs *L. danicus*
- *L. convexus* vs *L. hargravesii*
- *L. danicus* vs *L. hargravesii*

A first filtering procedure was performed; the transcripts with less than one read count in each sample were removed by using the edgeR package. The overall quality of the experiment was evaluated on the basis of the similarity among samples (now treated as replicates), by a PCA analysis using the filtered expression values of the transcripts. Using this approach, the detection of any samples with unexpected behavior was possible. Transcripts are considered significantly differentially expressed if the false discovery rate (FDR) of the statistical test was less than 0.05. No filter to the fold-change has been applied.

• **Transposable elements analysis.** A specific analysis on transposon elements was performed for the four species. In particular, annotation of the TE elements was performed with RepeatMasker and the information from the DE analysis and variant calling was linked to them.

The data produced from Sequentia in the analysis steps just described were further investigated and visualized. The visualization of the DE results was performed with the Venn diagram generator of the Microarray Center CRP-Sante (http://www.bioinformatics.lu/) and REVIGO (REduce and VIsualize Gene Ontology). The latter was used for the summary and visualization of the significant gene ontology terms (Supek et al., 2011). The resulting lists of GO terms were large and highly redundant and thus difficult to interpret. For that reason, REVIGO was selected since it can summarize long, unintelligible lists of GO terms by finding a representative subset of the terms using a simple clustering algorithm. In the end, the non-redundant GO term set is visualized to assist interpretation. The TreeMap visualization was selected in the current study (Fig. 4.2.1.).

Figure 4.2.1 The ''TreeMap'' view of REVIGO. Each rectangle is a single cluster representative. The representatives are joined into 'superclusters' of loosely related terms, visualized with different colours. Size of the rectangles may be adjusted to reflect either the p-value, or the frequency of the GO term in the underlying GOA database.

The expression values of the orthologous genes across all species were used for a hierarchical cluster analysis (HCA) and a canonical correlation analysis (CCA) with vegan R package. A corresponding heatmap was also created with Plotly (https://plot.ly).

Finally, the annotation results for each species were searched for terms related to temperature, stress, adaptation, circadian rhythm and transposable elements. The temperature and TE related transcripts that were found significantly different between temperatures in *L. aporus*, were blast searched for in the present dataset. In addition to these genes, specific *L. danicus* flagellate genes (Nanjappa et al., submitted) were blasted against all the four species assembled transcriptomes. A complete list of the genes blasted and their annotation is provided in Table 4.2.3.

Table 4.2.3 List of selected genes that were blasted against the *Leptocylindrus* transcriptomes. The genes were derived from the *L. aporus* transcriptomic analysis of Chapter 3 and the *L. danicus* transcriptomic analysis by Nanjappa et al. (submitted).

Genes Blasted (*L. aporus* 2015 and *L. danicus* 2012 transcript IDs)	Annotation
TR6847|c0_g1_i1	Heat Shock Factor protein 1 (HSF1)
TR1078|c1_g1_i1	Heat Stress transcription Factor C-1b (HSFC1b)
TR936|c0_g1_i1	Heat Stress transcription Factor B-2a (HSFB-2a)
TR264|c0_g1_i1	Heat Stress transcription Factor A-1a (HSFA-1a),
TR268|c0_g1_i1	Photosystem II 12 kDa extrinsic protein, chloroplastic (PSBU), stabilizes the structure of photosystem II oxygen-evolving complex (OEC), the ion environment of oxygen evolution and protects the OEC against heat-induced inactivation.
TR6964|c11_g1_i1	Ubiquinol Oxidase 4 Chloroplastic/Chromoplastic (AOX4), induced when plants are exposed to a variety of stresses including oxidative stress, chilling pathogen attack senescence and, in rice in particular, low temperature.
TR1252|c0_g1_i1	Stress-induced Yeast ortholog of the mammalian Mpv17 (SYM1)
TR6356|c2_g2_i1 TR7186|c6_g2_i10 TR6586|c2_g1_i1 TR6506|c2_g2_i3 TR6506|c2_g2_i5 TR7165|c11_g1_i1 TR6877|c1_g1_i1 TR7092|c10_g5_i1 TR6586|c2_g1_i1 TR6788|c3_g1_i1	Transposable Element (TE) related genes
MMETSP0321-20121206|1172	Tubulin-Tyrosine Ligase family protein (TTL), involved in the organisation of the neuronal microtubule network, in centriole stability, axoneme motility and mitosis.
MMETSP0321-20121206|1233	Intraflagellar transport protein 172 homolog (Ift172), required for the maintenance and formation of cilia.
MMETSP0321-20121206|4747	Cytoplasmic dynein 2 heavy chain 1 (DYNC2H1), possible motor for intraflagellar retrograde transport. Functions in cilia biogenesis.
MMETSP0321-20121206|7324	B9 domain-containing protein 1 (B9D1), component of a complex localized at the transition zone of primary cilia.
MMETSP0321-20121206|25141	Dynein heavy chain, axonemal (DNAH), force generating protein of respiratory cilia; involved in sperm motility and implicated in sperm flagellar assembly.
MMETSP0321-20121206|29637	Tetratricopeptide repeat protein 30a-like (TT30A), involved in intraflagellar transport.

4.3. Results

The phylogenetic tree based on ITS (Fig. 4.3.1) showed that all selected strains clustered according to species with a high bootstrap value while the intraspecific variability was low.

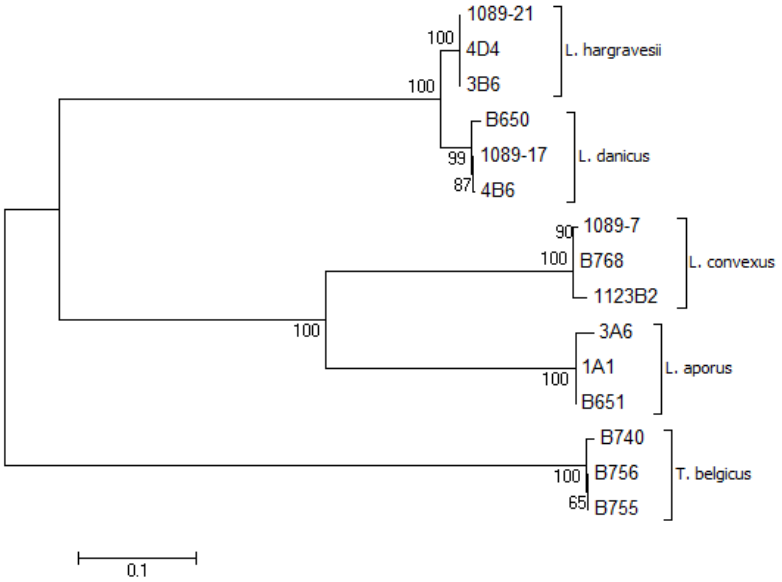
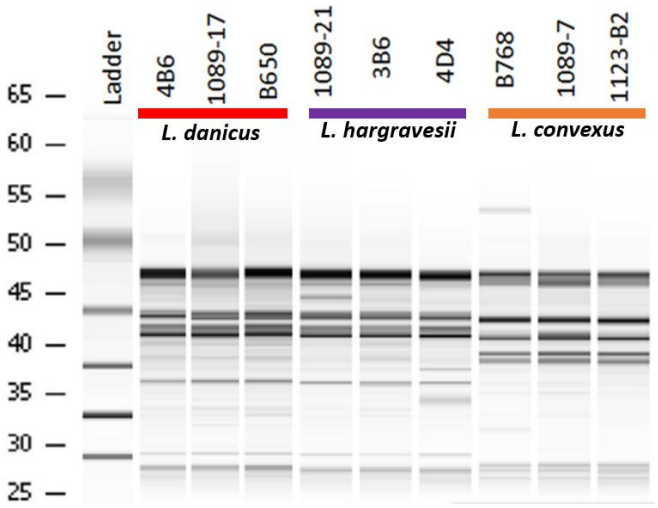


Figure 4.3.1 Maximum likelihood phylogenetic tree (Kimura-2 parameter model) based on ITS sequences of the selected *Leptocylindrus* strains for RNA sequencing. Numbers on branches represent bootstrap values (500 replicates). *Tenuicylindrus belgicus* clade serves as an outgroup.

As in Chapter 3, for the assessment of RNA quality the Bioanalyzer electrophoresis results and electropherograms were used and all samples were found of an acceptable quality for sequencing (Fig. 4.3.2). The *L. aporus* samples were the same as the ones at 19 °C in Chapter 3 and they were of good quality. They are not presented here again.



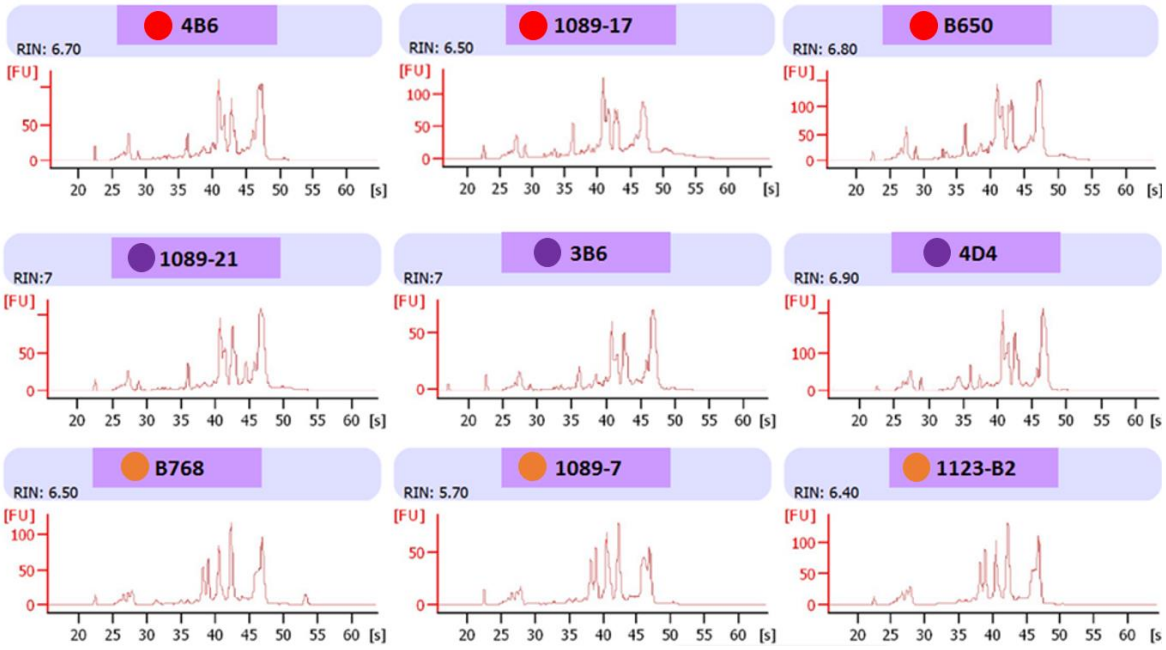


Figure 4.3.2 Bioanalyzer (electrophoresis above and electropherogram below) results of *L. danicus* (red), *L. hargravesii* (purple) and *L. convexus* (orange) RNA samples sent for sequencing.

Bioanalyzer results showed the typical pattern of a good quality diatom RNA with the three expected bands/ peaks except for *L. convexus,* where an extra double peak was obvious in all samples.

The quality check that was performed on the raw sequencing data established sequences of minimum length 35 bp and quality score 25 (Table 4.3.1).

Table 4.3.1 Resulting number of reads after the quality check for each species.

Sample Name	Species	N of reads before data quality control	N of reads after data quality control	N of reads after read processing
1A1	*L. aporus*	15,561,704	15,045,589	25,631,238
3A6	*L. aporus*	18,020,449	17,377,190	
B651	*L. aporus*	17,955,373	17,229,701	
1089-7	*L. convexus*	21,146,100	20,762,044	17,374,888
1123B2	*L. convexus*	21,492,174	21,194,391	
B768	*L. convexus*	19,621,719	19,318,073	
1089-17	*L. danicus*	18,700,334	18,401,071	21,897,354
4B6	*L. danicus*	21,134,252	20,305,593	
B650	*L. danicus*	23,033,307	21,356,367	
1089-21	*L. hargravesii*	17,242,020	16,923,027	16,725,603
3B6	*L. hargravesii*	20,907,779	20,560,987	
4D4	*L. hargravesii*	14,978,623	14,815,284	

Table 4.3.2 sums up the transcripts filtered at each step and presents statistics of the transcriptome assembly that were calculated using an assembly evaluation script.

Table 4.3.2 Number of transcripts produced after each filtration step and statistics on the final transcriptome for each species. N50 is the length of the longest contig in order for all contigs of at least that length to compose >50% of the assembly.

Transcripts	*L. aporus*	*L. convexus*	*L. danicus*	*L. hargravesii*
Raw	76,218	76,121	203,359	112,599
Clustering	33,728	20,254	33,809	27,173
Good quality	NA	19,210	32,219	25,814
Contaminants/ artefacts (removed)	294	332	413	1,450
Final curated assembly	***33,434***	***18,878***	***31,806***	***24,364***
Total length (bp)	36,347,313	22,589,874	27,558,193	22,864,803
%GC	41.16	38	44.63	44.84
Mean length (bp)	1087.14	1196.62	866.45	938.47
N50 (bp)	1,652	1,706	1,288	1,465

L. aporus and *L. danicus* have the largest transcriptome, followed by *L. hargravesii* and lastly *L. convexus.* Accordingly 28,632, 14,574, 24,915 and 19,316 protein sequences were translated with a minimum length of 50 aa for *L. aporus, L. convexus, L. danicus* and *L. hargravesii* respectively.

About 35.06% of *L. aporus* transcriptome was annotated. Similarly, 31.41% of *L. danicus,* 35.83% of *L. hargravesii* and 41.2% of *L. convexus* received any kind of annotation. These percentages are slightly lower than expected, which is typically between 50 – 80% in RNA-seq studies (Conesa et al., 2016) so a high proportion of the transcriptomes remained composed of unknown transcripts.

4.3.1. Variant calling analysis

The genetic polymorphism of each species and the related effect on gene products was explored through the variant calling analysis. The number of the raw variants for each species, as well as the number after the applied filters, is shown in the table and the figure below. SNPs were the prevailing variants for all species. *L. danicus* had at least three times more variants, mainly SNPs, compared to the other species (Table 4.3.1.1 and Fig. 4.3.1.1).

Table 4.3.1.1 Number of raw and filtered variants for each species.

Sample Name	Raw variants	Filtered variants
L. aporus	127,629	36,567
L. convexus	99,329	53,022
L. danicus	492,560	166,360
L. hargravesii	88,876	38,043

Figure 4.3.1.1 Variant calling results. The histograms represent the total number of filtered variants detected by SUPER for each of the studied species. The variants detected are Single Nucleotide Polymorphisms (SNPs), deletions (DEL) and insertions (INS).

In more detail for each species, the effects of the variants on the corresponding coding sequence are shown in table 4.3.1.2. In all species the majority of the variants led to synonymous and missense point mutations.

Table 4.3.1.2 Numbers of effect types and the corresponding putative impact of the variants in each species.

Effects on Gene Product	Impact	*L. danicus*	*L. aporus*	*L. convexus*	*L. hargravesii*
chromosome number variation	High	15	5	46	37
frameshift variant	High	87	108	266	276
frameshift variant & start lost	High	6	2	7	3
frameshift variant & stop gained	High	6	2	1	2
frameshift variant & stop lost	High	10	3	9	12
start lost	High	29	19	21	10
start lost & disruptive inframe deletion	High	0	0	0	1
start lost & disruptive inframe insertion	High	0	1	3	1
start lost & inframe deletion	High	2	0	1	0
start lost & inframe insertion	High	0	0	1	0
stop gained	High	269	87	175	126
stop gained & disruptive inframe deletion	High	0	0	0	1
stop gained & disruptive inframe insertion	High	0	0	0	1
stop gained & inframe insertion	High	0	0	2	2
stop lost	High	44	18	32	34
stop lost & disruptive inframe deletion	High	1	0	0	0
stop lost&inframe deletion	High	1	0	1	2
stop retained variant	High	159	28	25	34
disruptive inframe deletion	Moderate	254	19	159	223
disruptive inframe insertion	Moderate	211	10	97	165
inframe deletion	Moderate	120	22	76	87
inframe insertion	Moderate	98	22	44	95
missense variant	Moderate	43,248	10,957	19,058	9,374
5 prime UTR premature start codon gain variant	Low	1,144	602	567	477
initiator codon variant	Low	3	3	3	2
synonymous variant	Low	90,632	13,243	18,053	11,762
3 prime UTR variant	Modifier	19,761	6,693	7,594	8,406
5 prime UTR variant	Modifier	6,864	3,907	3,934	3,092
Sum		**162,964**	**35,751**	**50,175**	**34,225**

The number of high and moderate impact variants was calculated for each species strain (Fig 4.3.1.2). The majority of the moderate impact variants were missense variants which can be directly attributed to SNPs. *Leptocylindrus convexus* and *L. hargravesii* were the most uniform species regarding the variants' abundance while *L. danicus* and *L. aporus* had each a strain that possessed more variants than the other strains (4B6 and B651 respectively). Each class of moderate/high variant was calculated also as a percentage over the total number of variants in each species (Fig. 4.3.1.2). This made obvious that the species differed not only regarding the number of their variants but also regarding their composition.

Figure 4.3.1.2 Numbers of high and moderate impact variants (above) and their percentages over total variants (below) for each strain of *Leptocylindrus* species.

The GO terms enriched for those transcripts that were significantly affected by high impact variants were used for Venn diagrams among species and visualization with REVIGO (Fig. 4.3.1.3 - 7). 42.75% of *L. aporus*, 25.7% of *L. convexus*, 37.22% of *L. danicus* and 44.33% of *L. hargravesii* high impact variants received no annotation at all. 50.6% of the high impact variants shared the

same exact GO terms in all four species while the two species that greatly differentiated regarding the functions of the variants were *L. danicus* and *L. convexus.*

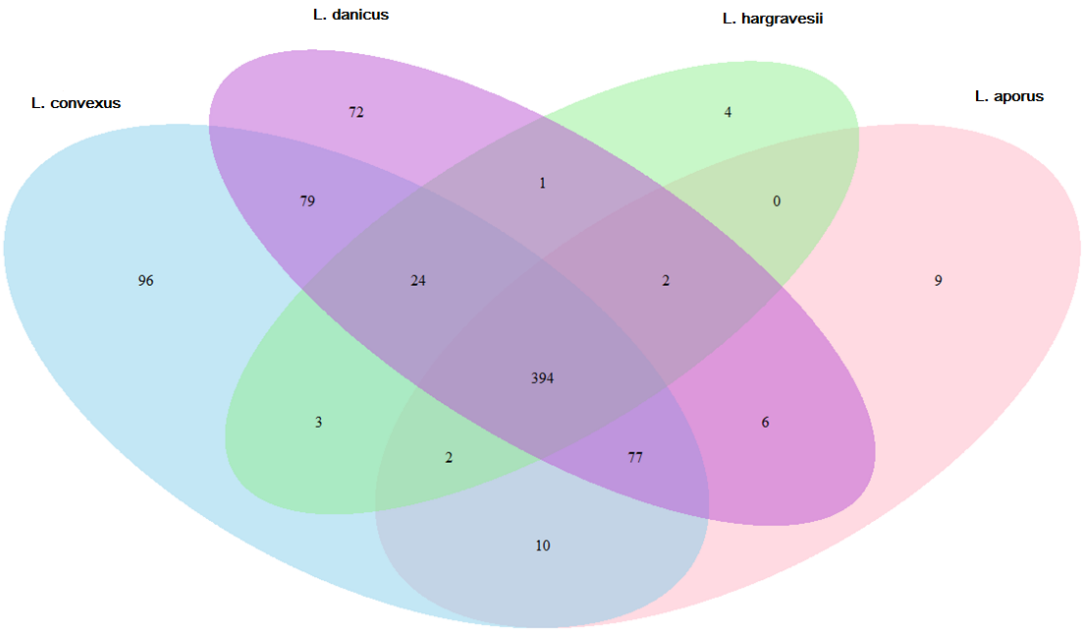


Figure 4.3.1.3 Venn diagram of the GO enriched terms of the transcripts that are significantly affected by high impact variants.

The annotations of the shared high impact variants revealed many functions that were equally affected by the observed polymorphism in all species, such as quite basic metabolic and biosynthesis GO terms but also post translational modifications and processes related to transportation of molecules (Fig. 4.3.1.4).

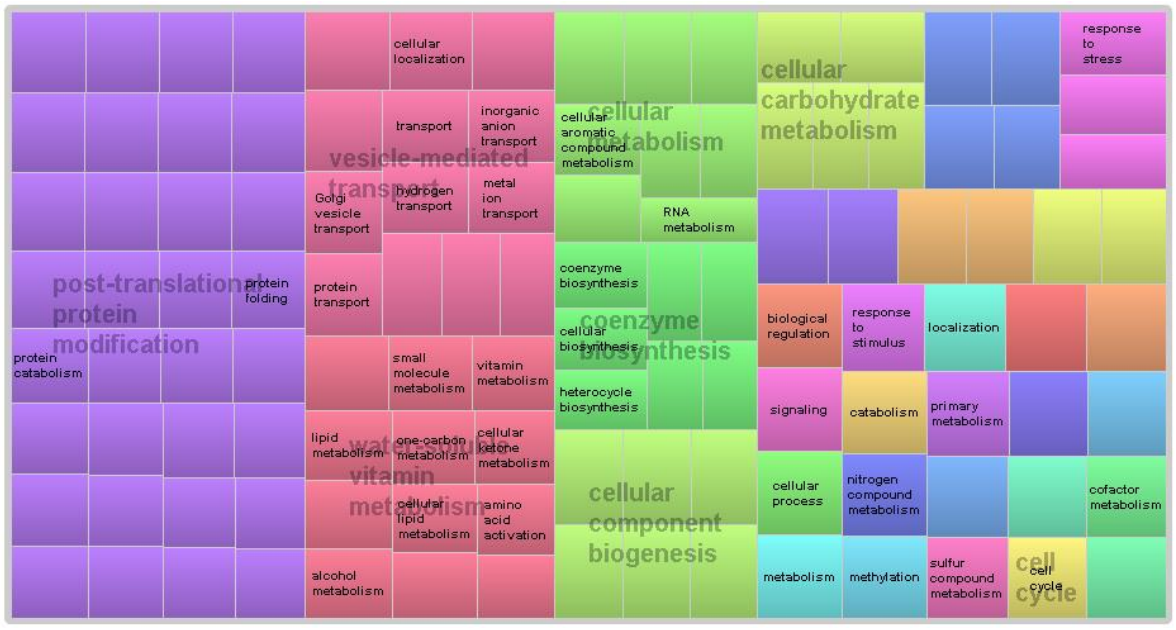


Figure 4.3.1.4 Biological process GO enrichment "TreeMap" view of REVIGO for the shared high impact variants in all four species. Size of the rectangles is adjusted to reflect the frequency of the GO term in the dataset.

Many of the *L. danicus* and *L. convexus* common biological processes (organelle fision, lipoprotein metabolism, protein localization to organelle, lipid modification) indicate as well an effect of variants on transportation of protein complexes or organelles for these two species (Fig.4.3.1.5).

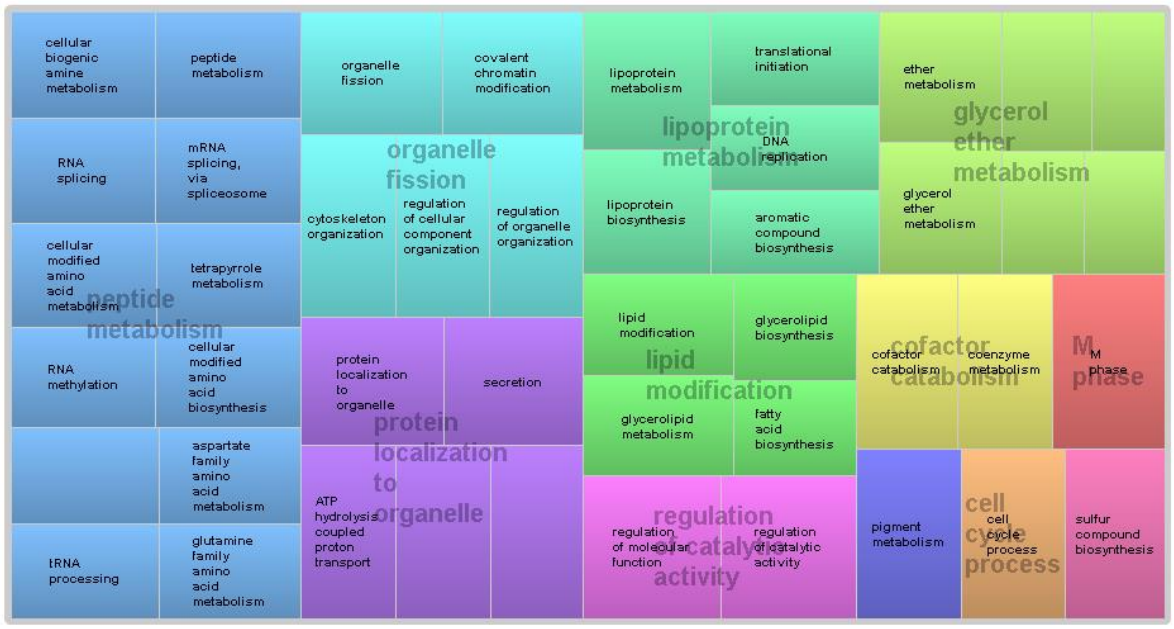


Figure 4.3.1.5 Biological process GO enrichment ''TreeMap'' view of REVIGO for shared high impact variants between *L. convexus* and *L. danicus*. Size of the rectangles is adjusted to reflect frequency of the GO term in the dataset.

Nevertheless, there were variants that affected functions unique to certain species. The *L. danicus* unique variants had a high impact again to transportation (vacuolar transport) but also more specifically to reactions and pathways related to stress (response to external stimulus, protein repair) (Fig.4.3.1.6). The pigment biosynthesis could be a result of the higher chloroplast abundance in this species compared to the others. However, *L. hargravesii* could also share variants affecting pigment biosynthesis due to their highly similar morphology, but the low annotation percentage might have influenced the lack of their detection.

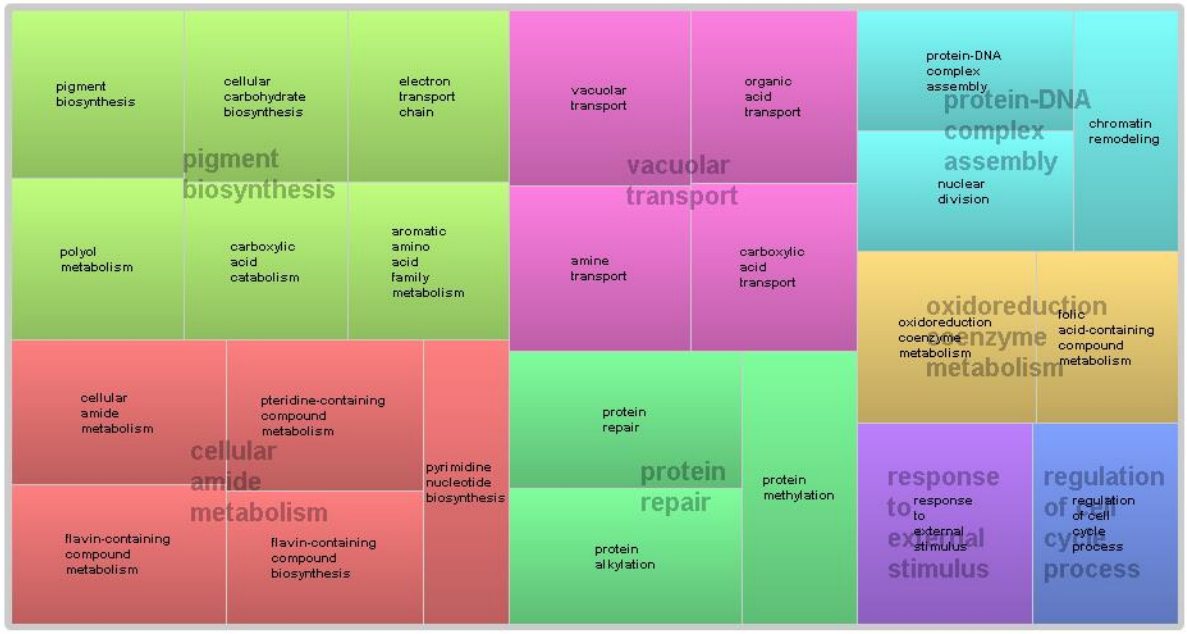


Figure 4.3.1.6 Biological process GO enrichment "TreeMap" view of REVIGO for *L. danicus* unique high impact variants. Size of the rectangles is adjusted to reflect the frequency of the GO term in the dataset.

The *L. convexus* variants had a high impact on many similar functions such as nucleocytoplasmic transport, autophagy and protein processing (Fig. 4.3.1.7).

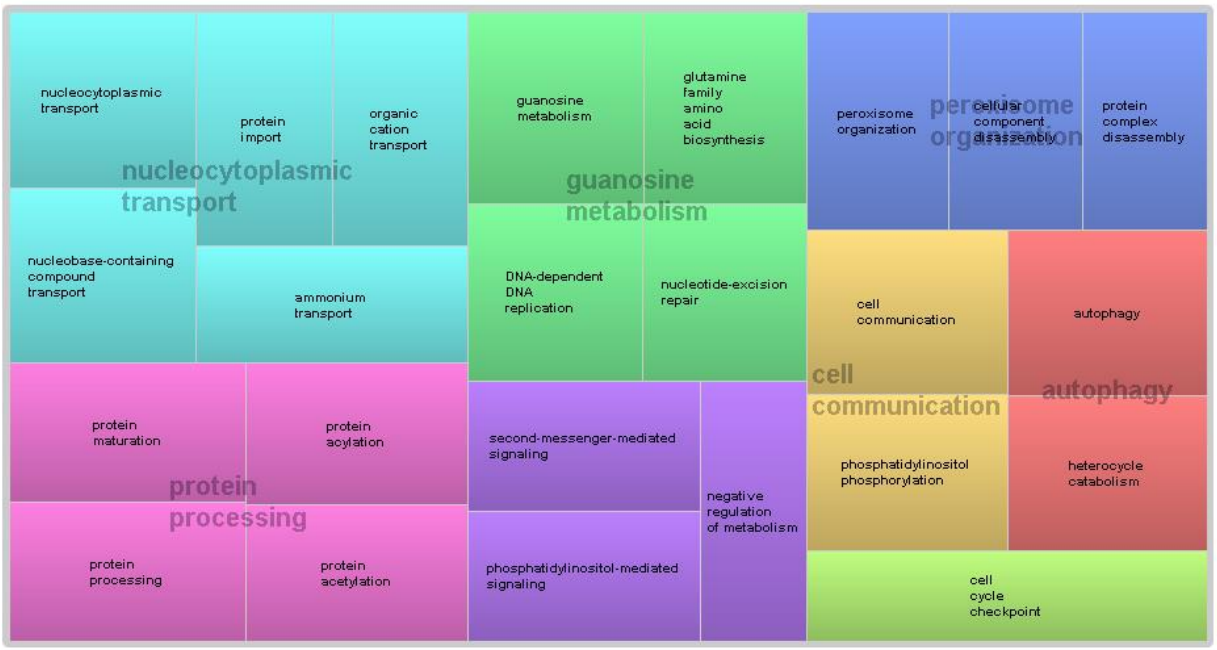


Figure 4.3.1.7 Biological process GO enrichment "TreeMap" view of REVIGO for *L. convexus* unique high impact variants. Size of the rectangles is adjusted to reflect the frequency of the GO term.

The unique high variants of *L. aporus* were related to aspartate family amino acid biosynthesis, whereas those of *L. hargravesii* were related to protein glycosylation.

4.3.2. Differential expression analysis among strains of each species

The differential expression analysis among the strains in each species produced lists of significantly up and down regulated transcripts for each pair (Fig. 4.3.2.1). The differences among the different strains within each species were mainly of the same intensity with the exception of 1089-17 - 4B6 pair in *L. danicus* and 1A1-3A6 in *L. aporus* that were more different than the rest.

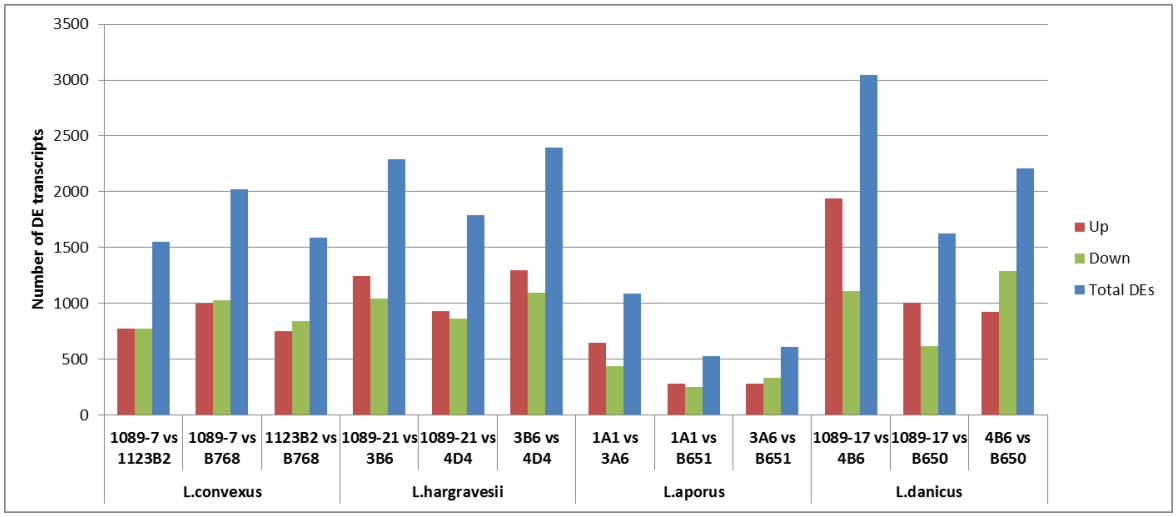
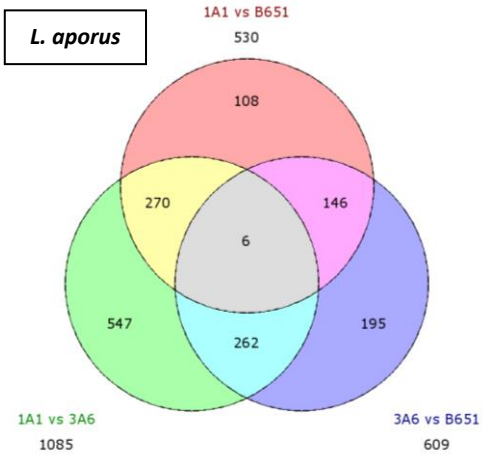
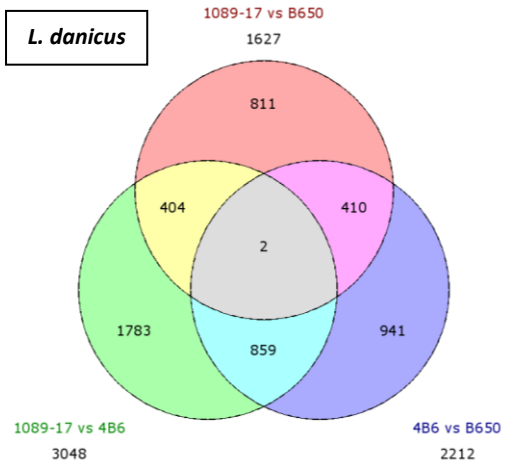


Figure 4.3.2.1 Number of significantly (FDR<0.05) up and downregulated transcripts produced by the NOIseq differential expression analysis in *L. convexus, L. hargravesii, L. aporus* and *L. danicus*.

The unique significantly different transcripts for each pair were also calculated and represented in the following Venn diagrams and table. Significant DE transcripts shared among all three pairs of strains in all species were very low, even zero in the case of *L. hargravesii.*





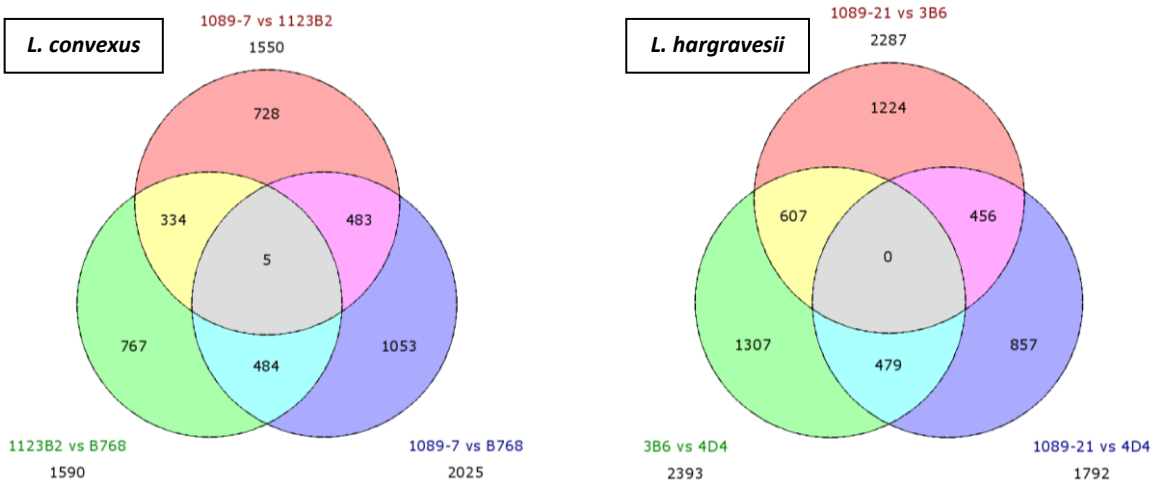


Figure 4.3.2.2 Venn diagrams of the significant DE transcripts in each species.

The unique transcripts were many more in the 1A1 – 3A6 pair compared to the rest *L. aporus* pairs and in 1089-17 – 4B6 pair in *L. danicus,* which was expected due to the higher number of significant DE transcripts detected in these pairs.

Table 4.3.2.1 Unique significant DE transcripts of the pairs of strains within each species.

Unique Significant Transcripts												
	L. convexus			*L. hargravesii*			*L. aporus*			*L. danicus*		
	1089-7 vs 1123B2	**1089-7 vs B768**	**1123B2 vs B768**	**1089-21 vs 3B6**	**1089-21 vs 4D4**	**3B6 vs 4D4**	**1A1 vs 3A6**	**1A1 vs B651**	**3A6 vs B651**	**1089-17 vs 4B6**	**1089-17 vs B650**	**4B6 vs B650**
Up	382	505	312	658	421	767	309	28	101	1150	582	425
Down	346	548	455	566	436	540	238	80	94	633	229	516
Total	728	1053	767	1224	857	1307	547	108	195	1783	811	941

The log2(FC) values of the significantly different transcripts ranged from -1.5 to -8 and 1.5 to 7 for *L. aporus*, from -2.5 to -10 and 2.6 to 10 for *L. danicus* and *L. convexus*, and from -2.8 to -14 and 2.5 to 14 for *L. hargravesii.* An arbitrary threshold at -6 and 6 respectively was set in order to detect the highly differentially expressed genes (Table 4.3.2.2). *L. aporus* showed significantly less high fold significant DE transcripts than any other species while within species the 1089-17 - 4B6 pair in *L. danicus* stuck out due to the many more high fold significant DE transcripts.

Table 4.3.2.2 High fold significant DE transcripts in each species.

	L. convexus			*L. hargravesii*			*L. aporus*			*L. danicus*		
	1089-7 vs 1123B2	**1089-7 vs B768**	**1123B2 vs B768**	**1089-21 vs 3B6**	**1089-21 vs 4D4**	**3B6 vs 4D4**	**1A1 vs 3A6**	**1A1 vs B651**	**3A6 vs B651**	**1089-17 vs 4B6**	**1089-17 vs B650**	**4B6 vs B650**
Up	88	91	83	187	150	146	5	1	2	149	119	87
Down	65	80	102	108	166	185	0	1	4	145	61	72
Total	153	171	185	295	316	331	5	2	6	294	180	159

L. hargravesii had the highest percentage of highly differentially expressed transcripts, *L. danicus* and *L. convexus* followed with a similar level (with the exception of 1089-17 – 4B6 pair), and *L. aporus* with a significantly lower percentage (Fig. 4.3.2.3).

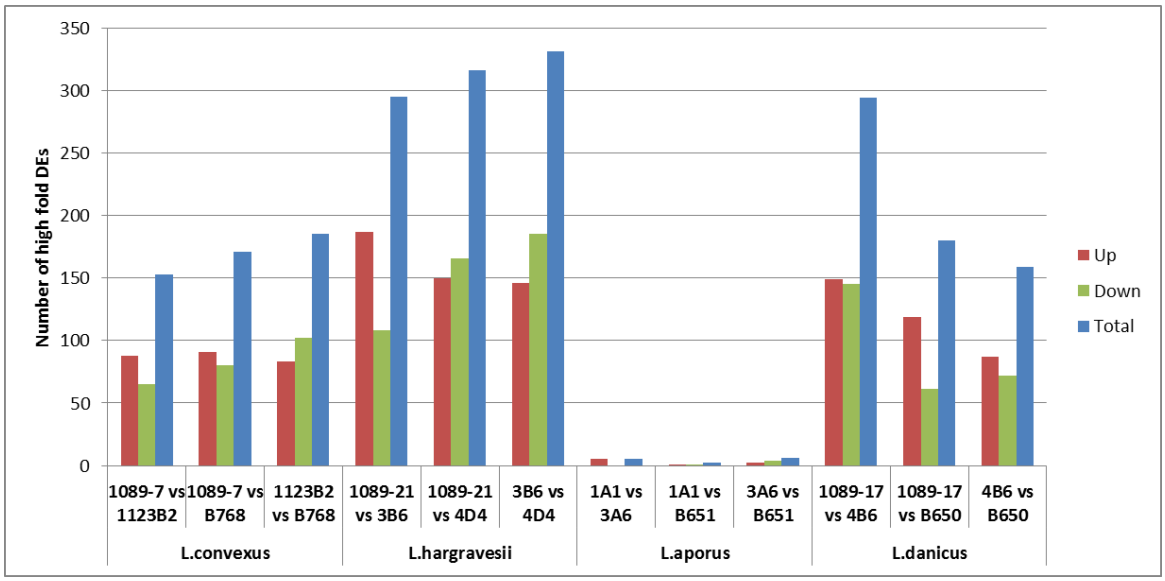


Figure 4.3.2.3 Barplot of high fold differentially expressed transcripts in each species.

For each species pair, the percentages of the significant DE transcripts that did not receive any annotation at all (annotated as "no description") are presented in Figure 4.3.2.4. In *L. danicus* the significant DE transcripts with no annotation reached up to almost 90% in the 4B6 – B650 pair while the rest pairs of all species varied between 35 – 60%.

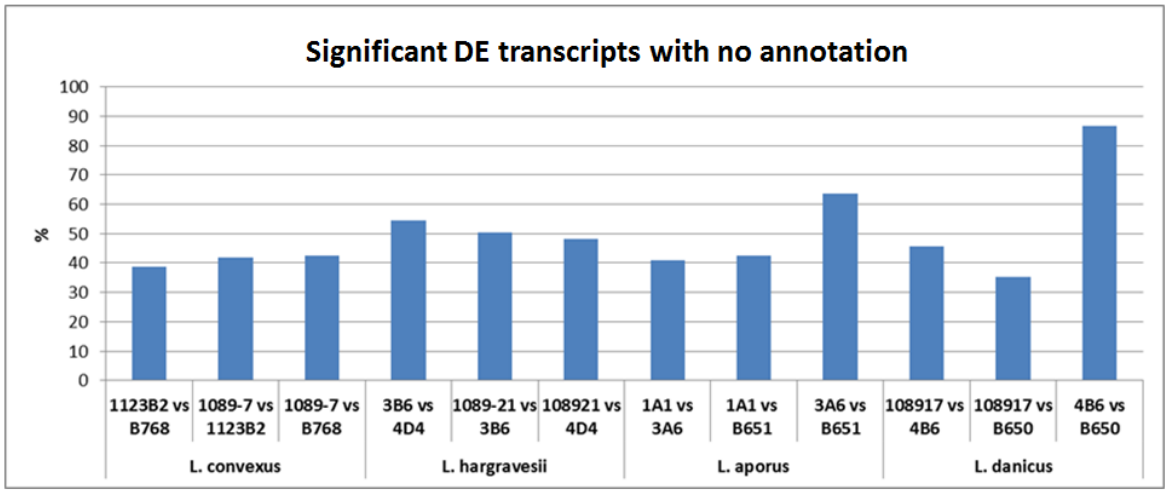


Figure 4.3.2.4 Percentage of significant DE transcripts that did not receive any annotation in each species comparison pair.

The significantly enriched GO terms of the significant DE genes among strains indicated the functions for which each pair of strains mainly differed.

GO terms that were represented by small rectangles in the *L. aporus* 1A1-3A6 pair were exocytosis, aminoacid transmembrane transport, chromatin modification, oxidation-reduction process (Fig. 4.3.2.5). Within the target of paramycin (TOR) signaling, which regulates cell growth and metabolism in response to environmental cues, signal transduction and DNA integration were also included. Functions related to transportation (exocytosis, aminoacid transmembrane transport, signal transduction, purine nucleobase transport) but also cell division (cytokinesis, kinetochrome assembly and chromatin modification) were present in all *L. aporus* pairs (Fig. 4.3.2.5 - 7).

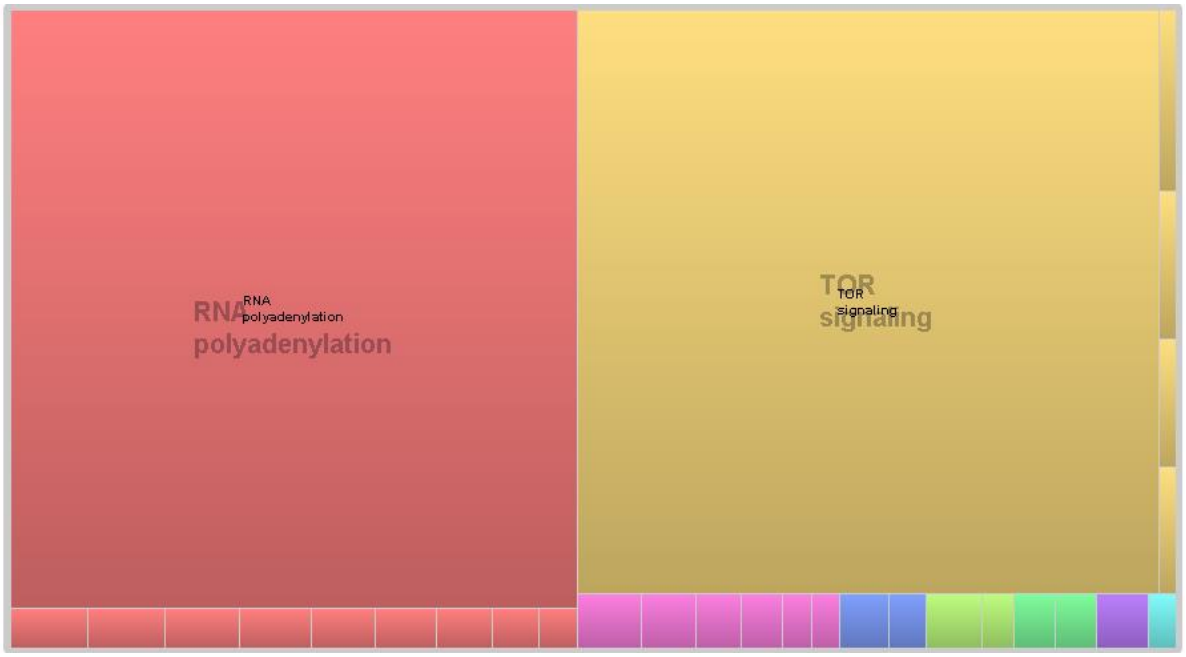


Figure 4.3.2.5 Biological process GO enrichment (FDR <=0.05) ''TreeMap'' view of REVIGO for *L. aporus* 1A1 vs 3A6. Size of the rectangles is adjusted to reflect the GOEA FDR corrected p-value.

The 1A1-B651 pair differed in several metabolic processes which shows that within a species one strain might show a higher or lower standard energy production than another.

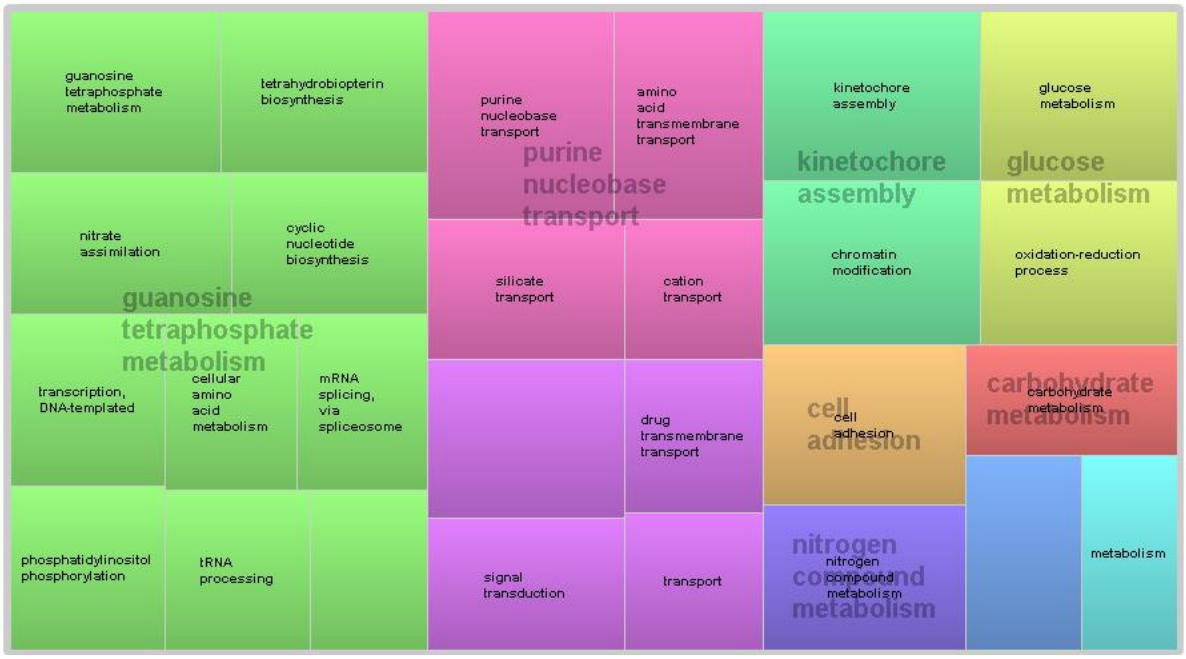


Figure 4.3.2.6 Biological process GO enrichment (FDR <=0.05) ''TreeMap'' view of REVIGO for *L. aporus* 1A1 vs B651. Size of the rectangles is adjusted to reflect the GOEA FDR corrected p-value.

GO terms that were represented by small rectangles in the 3A6-B651 pair were protein insertion into membrane and transmembrane transport. DNA integration was part of the RNA polyadenylation rectangle.

Figure 4.3.2.7. Biological process GO enrichment (FDR <=0.05) ''TreeMap'' view of REVIGO for *L. aporus* 3A6 vs B651. Size of the rectangles is adjusted to reflect the GOEA FDR corrected p-value.

For *L. danicus* pairs, (Fig. 4.3.2.8 - 10), there were again some transport related functions enriched in the comparison among strains (nucleoside transmembrane transport, intra-Golgi vesicle mediated transport) but the metabolic processes dominated in the case of this species.

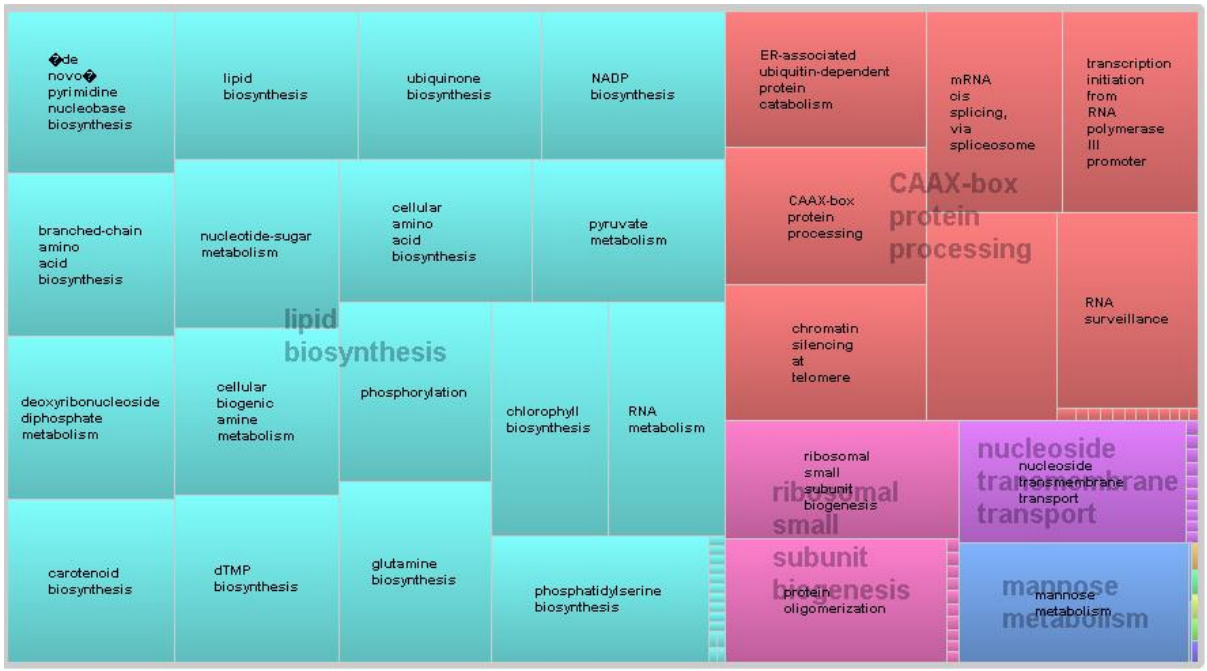


Figure 4.3.2.8 Biological process GO enrichment (FDR <=0.05) ''TreeMap'' view of REVIGO for *L. danicus* 1089-17 vs 4B6. Size of the rectangles is adjusted to reflect the GOEA FDR corrected p-value.

Although the parent GO terms were different among *L. danicus* strain pairs, several child terms like chlorophyll biosynthesis, RNA metabolism, RNA surveillance were similar, due to the fact that a child term can belong to more than one parent GO term. Transcription regulation terms were more enriched in the pair of 1089-17 and B650 (negative regulation of transcription from RNA polymerase III promoter, histone H3-K79 methylation) and then in the pair B650 and 4B6 (nuclear-transcribed mRNA catabolism, no-go decay).

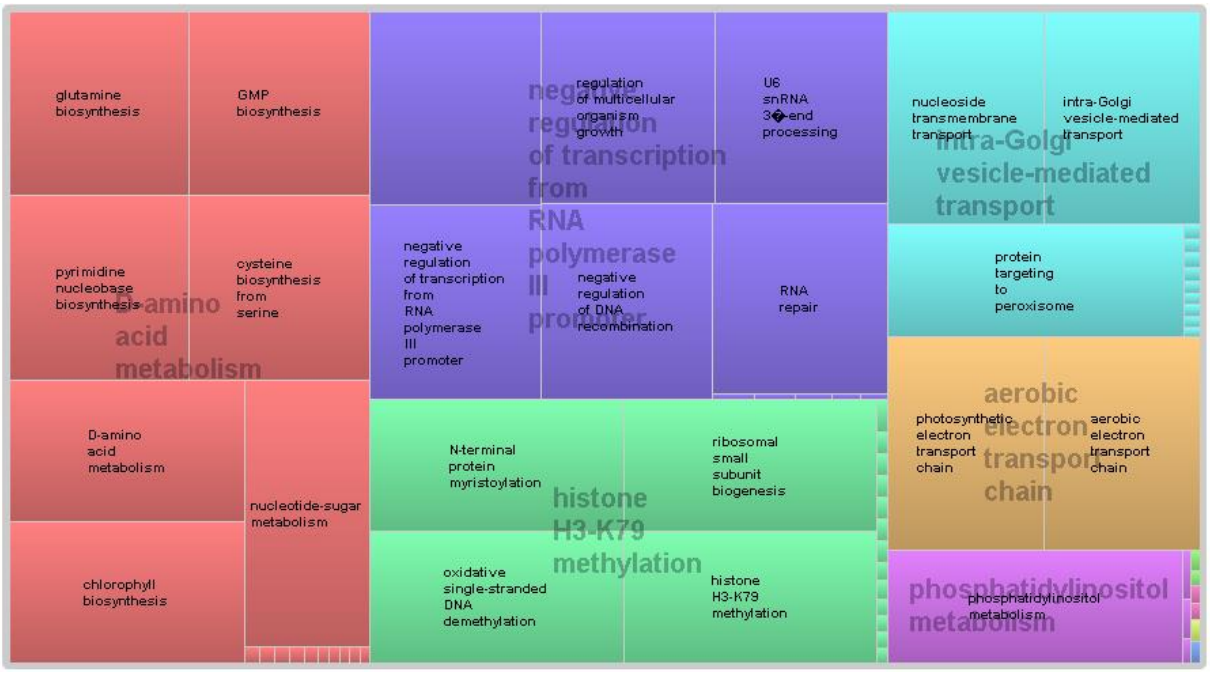


Figure 4.3.2.9 Biological process GO enrichment (FDR <=0.05) ''TreeMap'' view of REVIGO for *L. danicus* 1089-17 vs B650. Size of the rectangles is adjusted to reflect the GOEA FDR corrected p-value.

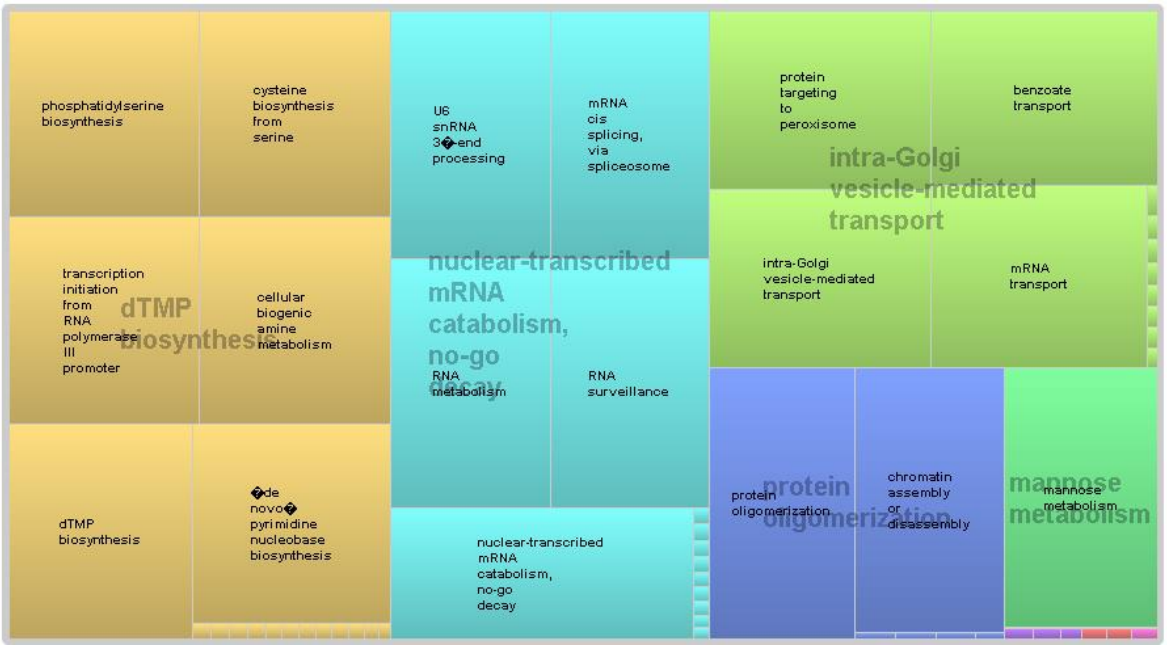


Figure 4.3.2.10 Biological process GO enrichment (FDR <=0.05) ''TreeMap'' view of REVIGO for *L. danicus* 4B6 vs B650. Size of the rectangles is adjusted to reflect the GOEA FDR corrected p-value.

In *L. convexus* (Fig. 4.3.2.11 - 13), cell division functions (DNA replication, regulation of microtubule polymerization, chromosome segregation, mitotic anaphase) were enriched in the 1123B2 – B768 pair and then also present but less enriched in the 1089-7 – 1123B2 pair, and even less in the third pair B768 – 1089-7.

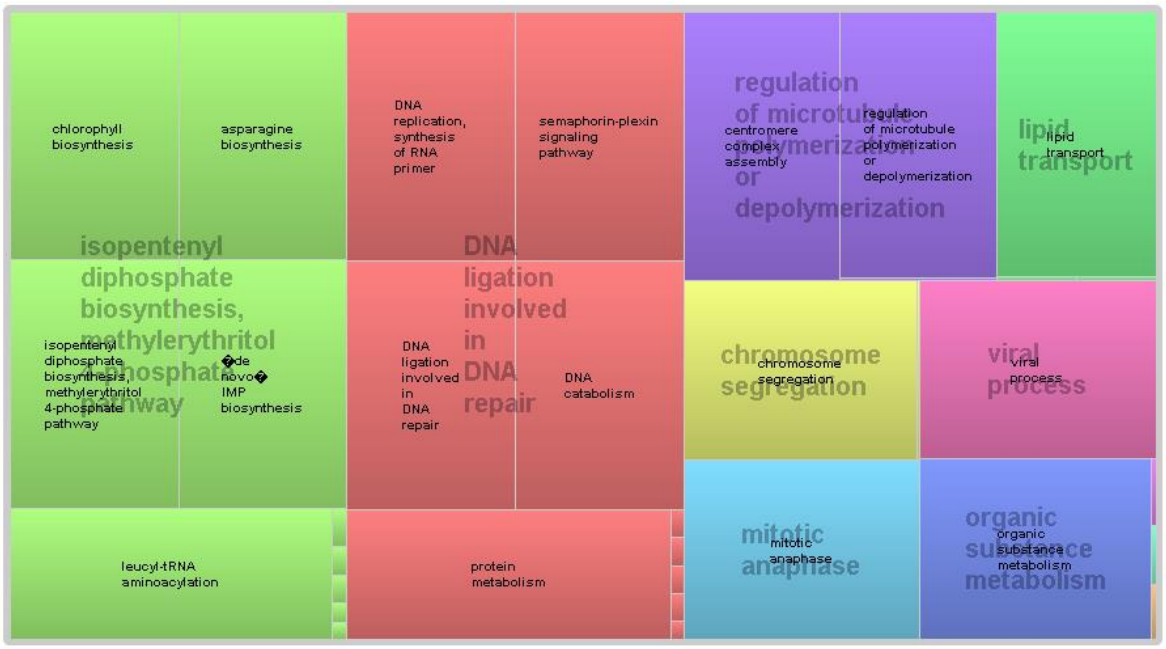


Figure 4.3.2.11 Biological process GO enrichment (FDR <=0.05) ''TreeMap'' view of REVIGO for *L. convexus* 1123B2 vs B768. Size of the rectangles is adjusted to reflect the GOEA FDR corrected p-value.

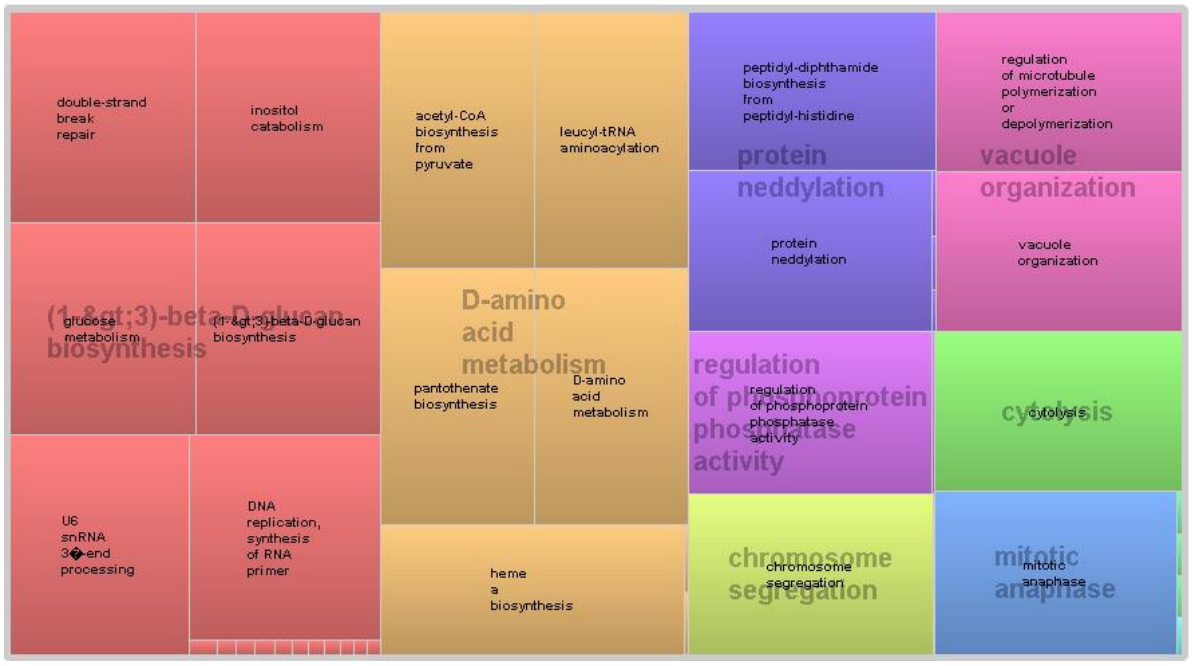


Figure 4.3.2.12 Biological process GO enrichment (FDR <=0.05) ''TreeMap'' view of REVIGO for *L. convexus* 1089-7 vs 1123B2. Size of the rectangles is adjusted to reflect the GOEA FDR corrected p-value.

Figure 4.3.2.13 Biological process GO enrichment (FDR <=0.05) ''TreeMap'' view of REVIGO for *L. convexus* 1089-7 vs B768. Size of the rectangles is adjusted to reflect the GOEA FDR corrected p-value.

In all three *L. hargravesii* pairs (Fig. 4.3.2.14 – 16) the biggest difference was related to lipid, peptide or chlorophyll biosynthesis.

Figure 4.3.2.14 Biological process GO enrichment (FDR <=0.05) ''TreeMap'' view of REVIGO for *L. hargravesii* 3B6 vs 4D4. Size of the rectangles is adjusted to reflect the GOEA FDR corrected p-value.

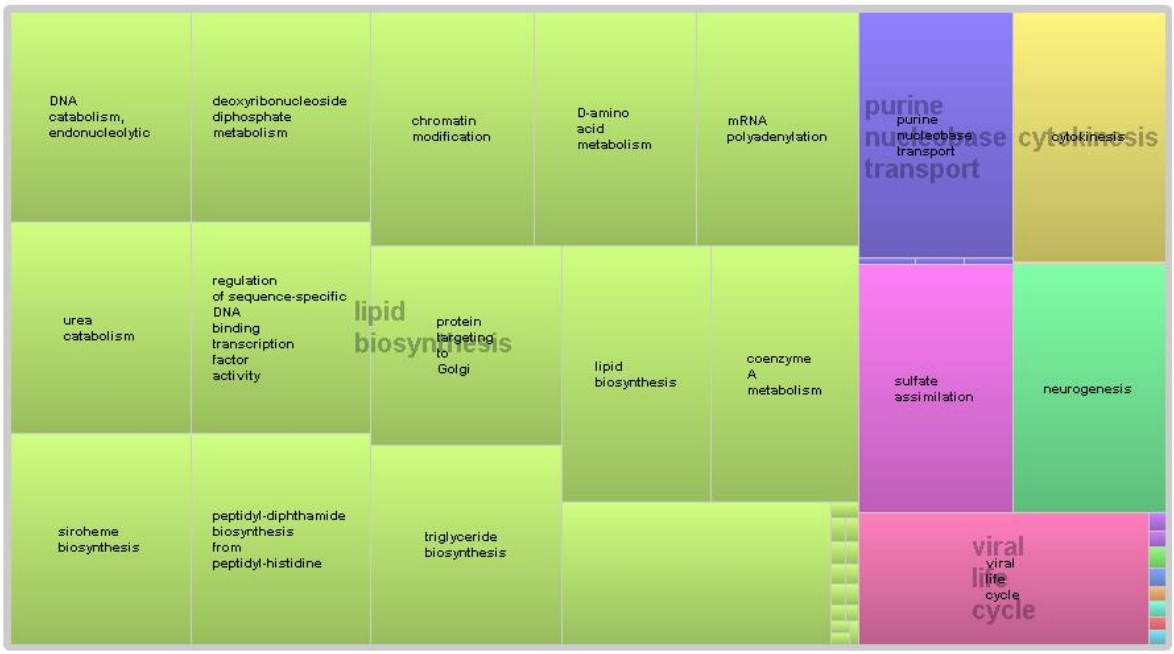


Figure 4.3.2.15 Biological process GO enrichment (FDR <=0.05) "TreeMap" view of REVIGO for *L. hargravesii* 1089-21 vs 3B6. Size of the rectangles is adjusted to reflect the GOEA FDR corrected p-value.

Figure 4.3.2.16 Biological process GO enrichment (FDR <=0.05) "TreeMap" view of REVIGO for *L. hargravesii* 1089-21 vs 4D4. Size of the rectangles is adjusted to reflect the GOEA FDR corrected p-value.

4.3.3. Orthologous genes in *Leptocylindrus* species

The orthology analysis produced a list of orthologous genes which was more or less of the same number in all possible trio combinations but was higher when *L. danicus* and *L. hargravesii* were involved which was quite reasonable considering that they are genetically closer. Indeed, the orthologous genes analysis between species pairs revealed 3,900 – 5,000 genes shared by all

species pairs except for *L. danicus* and *L. hargravesii* that shared 12,756 orthologous genes, which is at least double the amount of the other pairs.

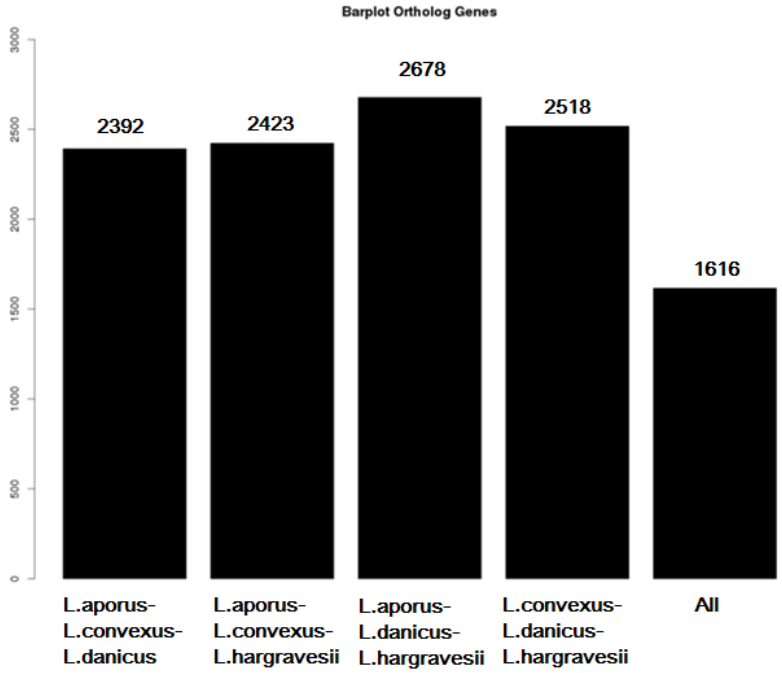


Figure 4.3.3.1 Numbers of orthologous genes found among groups of three species and all species.

The phylogenetic tree that was produced based on the sequences of the orthologous genes grouped *L. danicus* and *L. hargravesii* together, as expected based on rDNA-based phylogeny, whereas *L. aporus* that should be grouped with *L. convexus* was the most distantly related species (Fig.4.3.3.2).

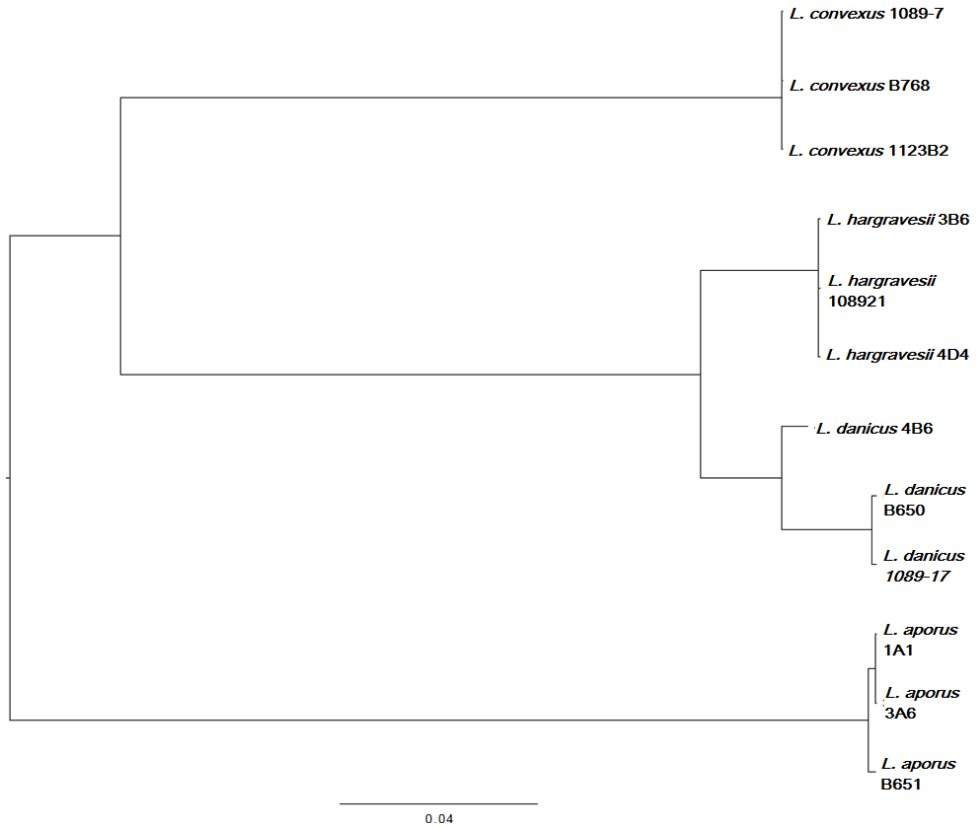
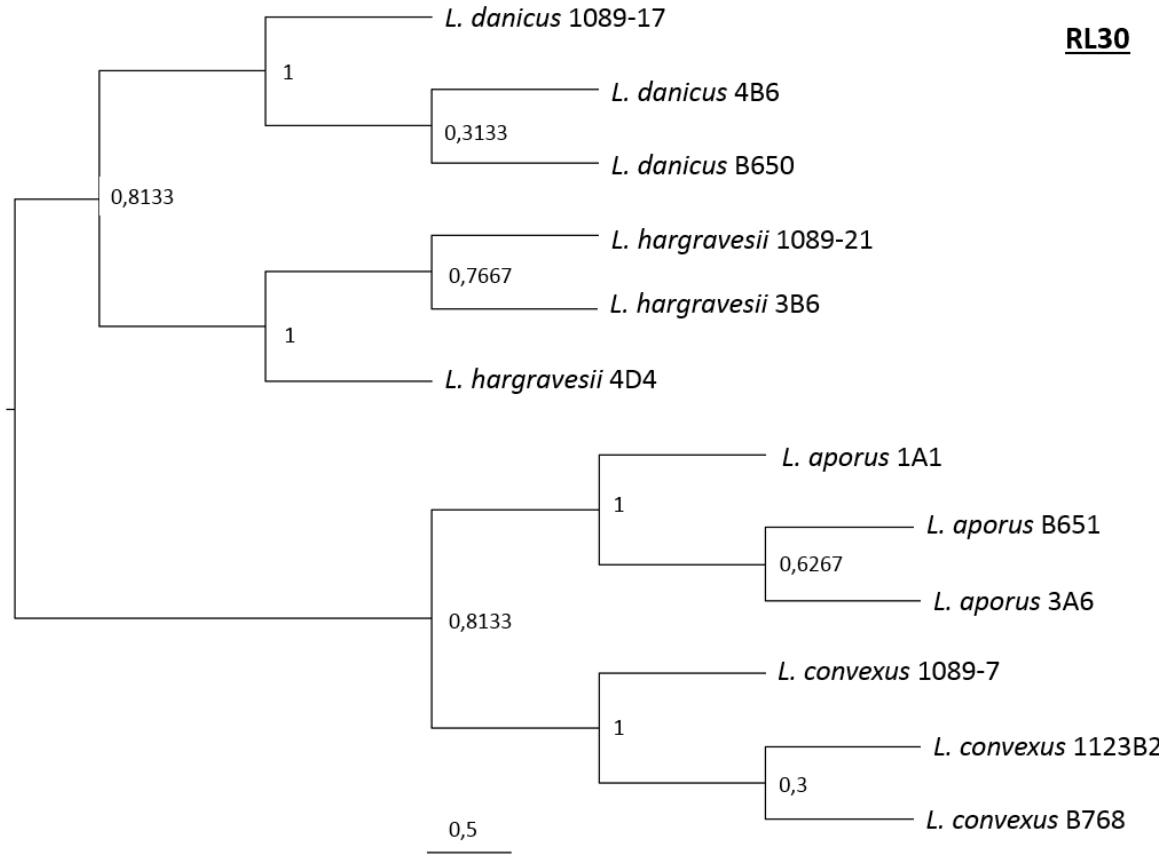
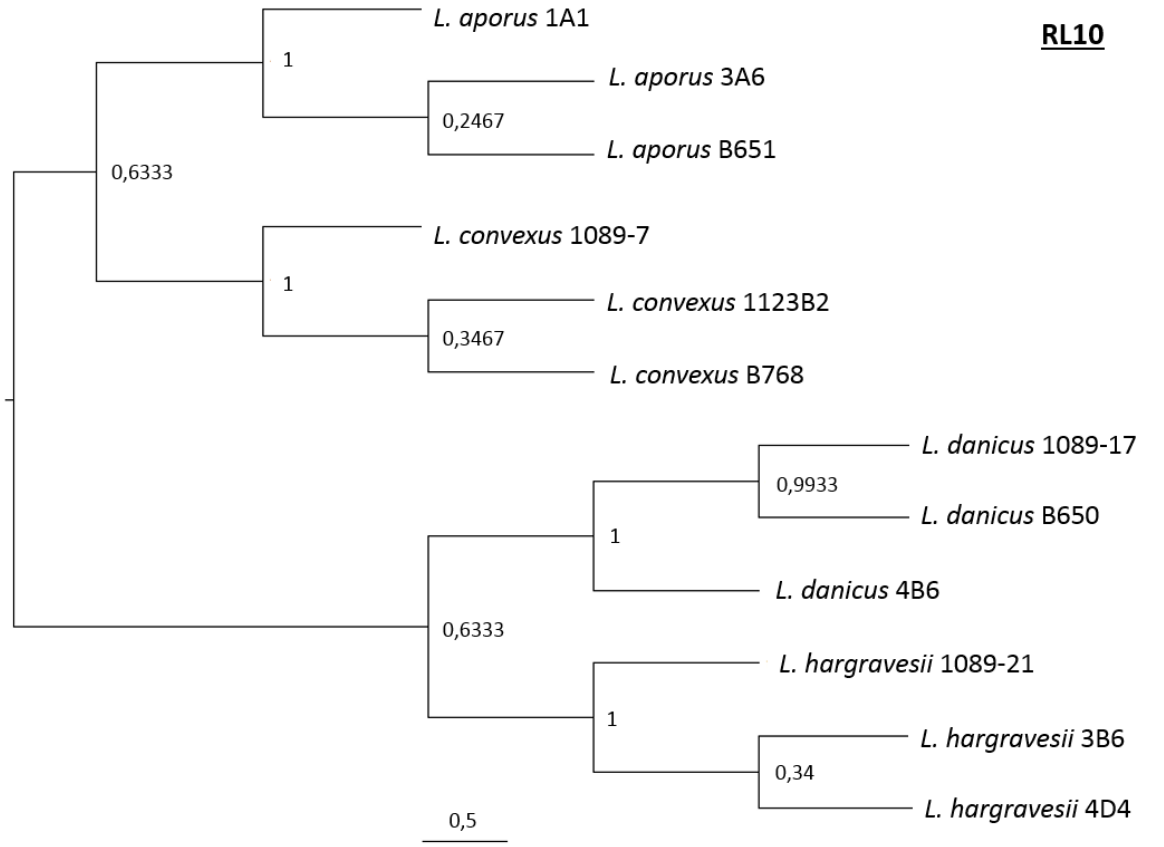


Figure 4.3.3.2 Neighbor-joining tree based on the orthologous sequences among species. The method used was alignment free calculation tree followed by phylogenetic reconstruction with PHYML tool. No bootstrap or any other similar statistics is supported by this free-alignment method.

Because of this unexpected result, more phylogenetic trees were constructed with maximum likelihood based on the alignment of five selected ribosomal proteins; ribosomal protein 30 60S large ribosomal subunit (RL30), ribosomal protein 7Ae 60S large ribosomal subunit (L7Ae), ribosomal protein 10 60S large ribosomal subunit (RL10), ribosomal protein L1, 60S ribosomal protein L31. Four out of the five gave the expected topology (Fig. 4.3.3.3), but L7Ae grouped *L. aporus* with *L. danicus* and *L. hargravesii* with *L. convexus* (not shown).


RL30
L. danicus 1089-17
1
L. danicus 4B6
0,3133
L. danicus B650
0,8133
L. hargravesii 1089-21
0,7667
L. hargravesii 3B6
1
L. hargravesii 4D4
L. aporus 1A1
1
L. aporus B651
0,6267
L. aporus 3A6
0,8133
L. convexus 1089-7
1
L. convexus 1123B2
0,3
L. convexus B768
0,5


RL10
L. aporus 1A1
1
L. aporus 3A6
0,2467
L. aporus B651
0,6333
L. convexus 1089-7
1
L. convexus 1123B2
0,3467
L. convexus B768
L. danicus 1089-17
0,9933
L. danicus B650
1
L. danicus 4B6
0,6333
L. hargravesii 1089-21
1
L. hargravesii 3B6
0,34
L. hargravesii 4D4
0,5

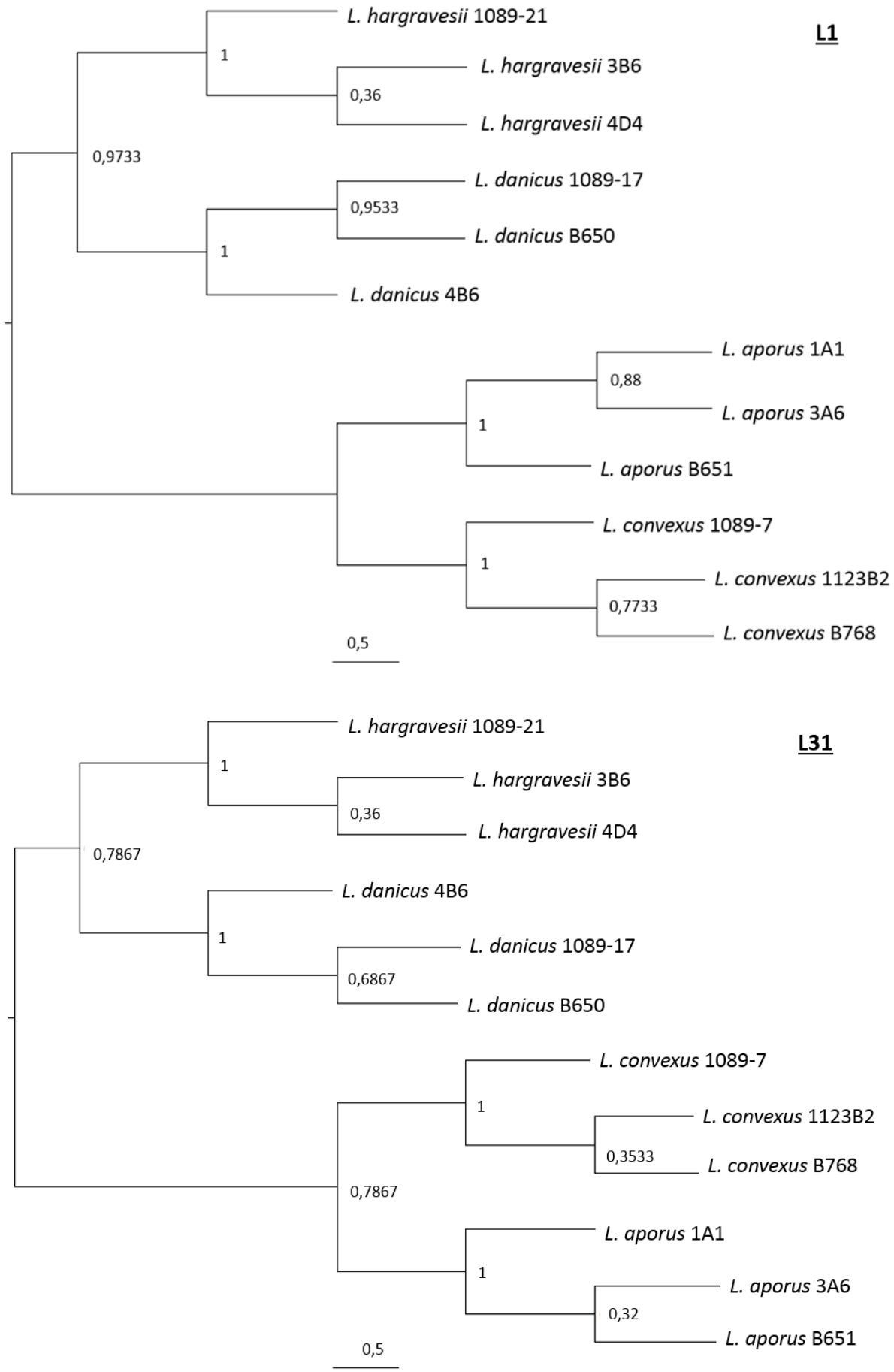


Figure 4.3.3.3 Maximum likelihood tree (Tamura, 500 bootstraps) based on the alignment of four selected ribosomal genes found orthologous among species, RL30, RL10, L1 and 60s ribosomal protein L31. The method used was alignment free calculation tree followed by phylogenetic reconstruction with PHYML tool.

Besides the sequences, the expression values of the orthologous genes were utilised in order to explore the similarity among strains and species. To this end, hierarchical clustering (HCA), principal component (PCA) and canonical correlation analysis (CCA) were used. All these methods were chosen since each shows a different aspect of the variance of data. The goal of HCA is to partition the objects into homogeneous groups, such that the within-group similarities are large compared to the between-group similarities. The principal components, on the other hand, are extracted to represent the patterns encoding the highest variance in the data set, whereas CCA maximizes the correlation between species scores and sample scores, constrained to be linear combinations of explanatory variables; neither PCA nor CCA maximize the separation between groups of samples directly.

The PCA analysis based on the expression values (filtered read counts) of the transcripts showed that all samples behaved as expected (Fig. 4.3.3.4), grouping according to species but also confirming the great distance of 4B6 from the rest of the *L. danicus* strains. In addition, the 3B6 *L. hargravesii* strain was not so significantly different from *L. convexus* and *L. danicus* while *L. aporus* strains seemed remarkably similar. Yet, this closeness of *L. aporus* strains in the PCA could be a result of a large difference from the rest of the strains of the other species rather than of the intraspecific similarity. The PCA results do not seem to follow our assumption that the expression patterns will be formed based on species seasonality but still the separation among species was possible.

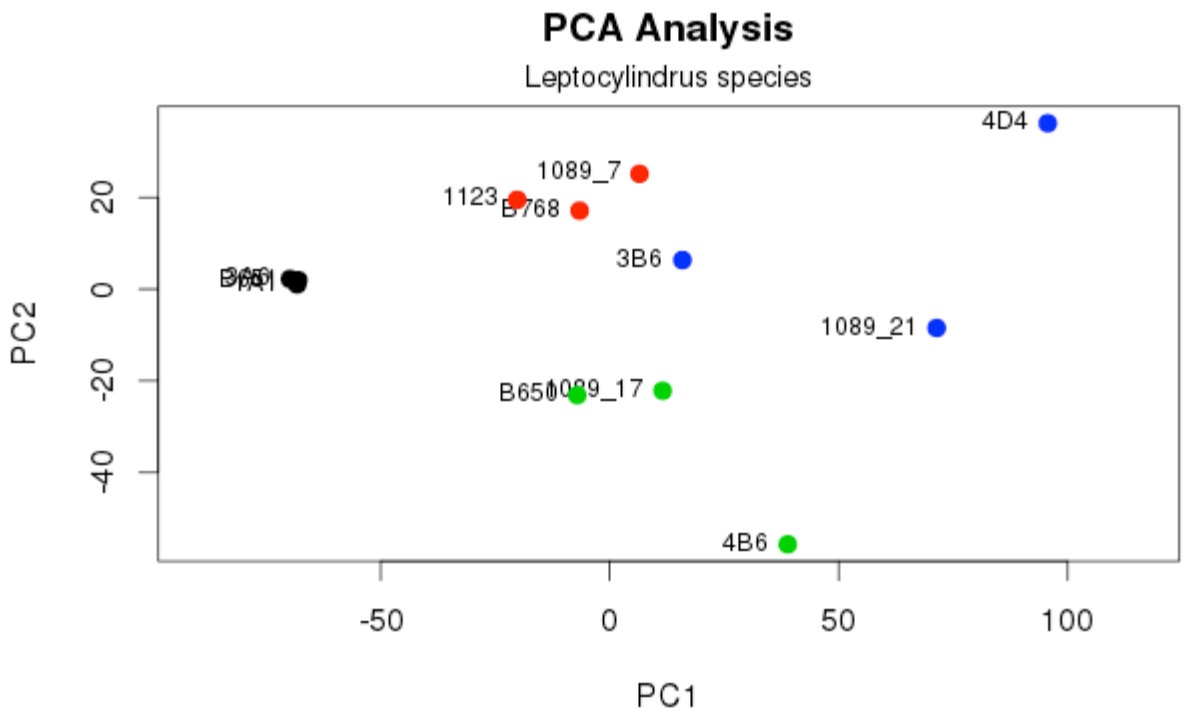


Figure 4.3.3.4 PCA analysis on filtered expression values of orthologous genes. Black: *L. aporus*, green: *L. danicus*, blue: *L. hargravesii*, red: *L. convexus.*

The expression values of the orthologous genes across all species were grouped according to species also in the HCA and CCA analysis (Fig. 4.3.3.5 - 6). In many high-dimensional real-world data sets, the most dominant patterns, i.e. those captured by the first principal components, are those separating different subgroups of the samples from each other. In these cases, the results from PCA and HCA would be similar. The HCA dendrogram is often represented together with a heatmap, which can help to identify the variables that appear to be characteristic for each sample cluster and be compared to CCA and PCA, where the synchronized variable representation provides the variables that are most closely linked to any groups emerging in the sample representation. The large distance of *L. aporus* from the rest of the species mentioned above was better represented in the cluster dendrogram and the corresponding heatmap where the *L. aporus* cluster was placed much further than the rest, but B651 was still different within the species (Fig. 4.3.3.5). 3B6 was the *L. hargravesii* strain that stuck out from the rest but in *L. danicus* 1089-17, not 4B6, was different. The heatmap though gave the same impression as the PCA analysis.

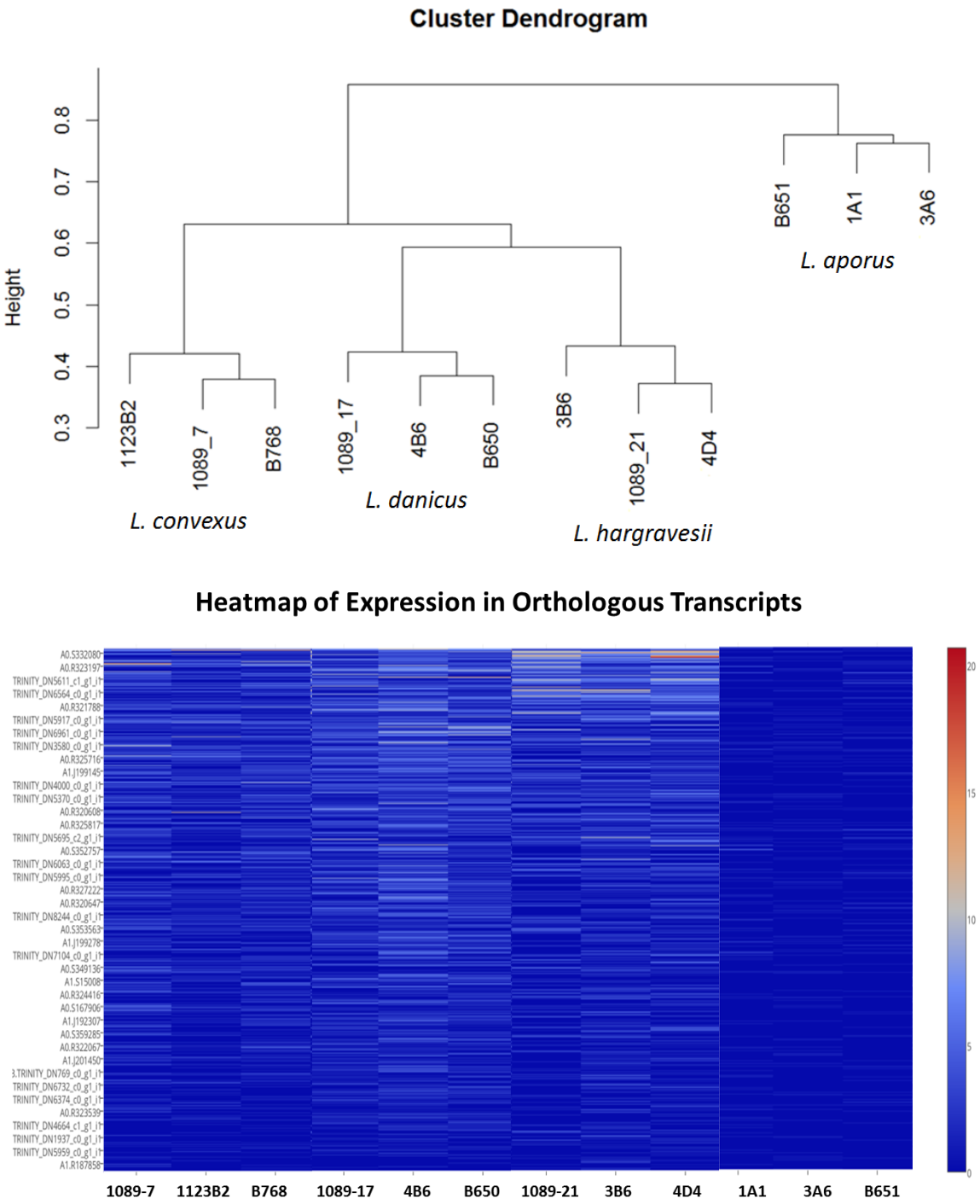


Figure 4.3.3.5 Hierarchical clustering of orthologous genes across all four species and corresponding expression heatmap.

The heatmap depicts the observed data without any pre-processing whereas PCA and CCA represent the data set in only a few dimensions, with some of the information in the data filtered out in the process. The discarded information is associated with the weakest signals and the least correlated variables in the data set, and it can often be safely assumed that much of it corresponds to measurement errors and noise. This makes the patterns revealed using PCA and CCA cleaner and easier to interpret than those seen in the heatmap, though taking the risk of

excluding weak but important patterns. For this reason, all methods were selected to explore the similarity of *Leptocylindrus* strains. Another difference is that HCA will always calculate clusters, even if there is no strong signal in the data, in contrast to the other two methods, which will present a plot similar to a cloud with samples evenly distributed. CCA represented the species better than PCA, showing almost equal distances among strains (Fig. 4.3.3.6). In this analysis, *L. convexus* was the species that showed the greatest distance from the rest; *L. hargravesii* was also quite far from the rest of the species while *L. danicus* and *L. aporus* were closer. The CCA result was the one that better depicted the separation of species based on their seasonality, placing the year-around *L. danicus* and *L. aporus* closer and the two species with a more specific seasonal distribution far apart.

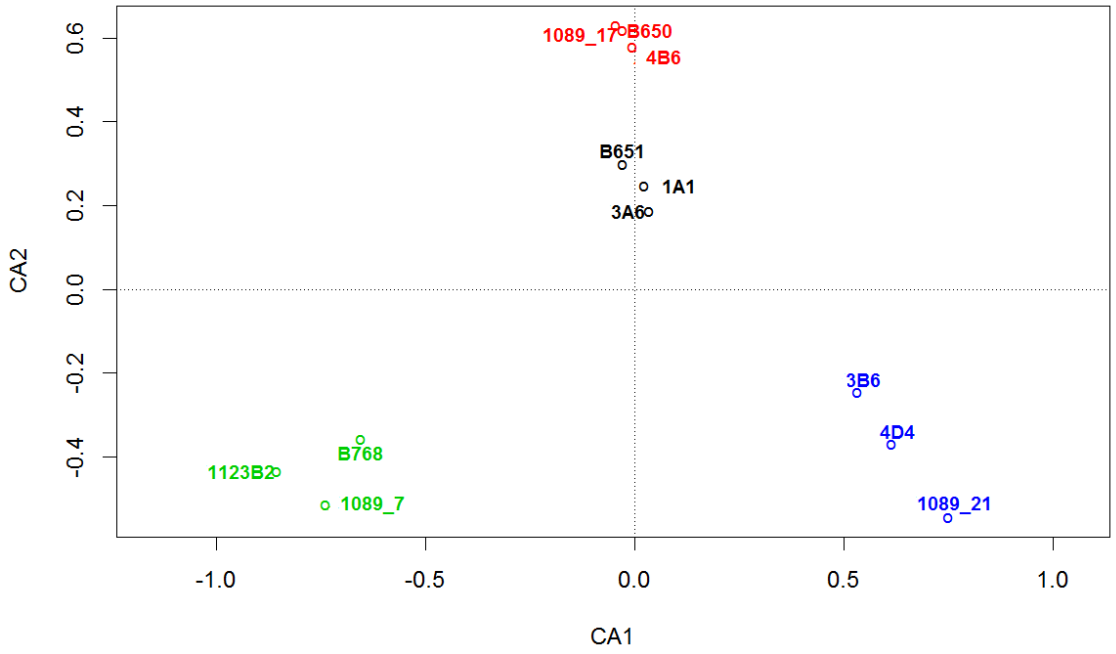


Figure 4.3.3.6 CCA plot of orthologous genes across all four species. Green: *L. convexus,* red: *L. danicus,* black: *L. aporus,* blue: *L. hargravesii.*

4.3.4. Differential expression analysis among *Leptocylindrus* species

Having kept all the strains in identical conditions allowed to perform a differential expression (DE) analysis on the genes found orthologous across species. The results to be presented from here on was limited to the transcripts found conserved across all four species (1616 transcripts) (Fig. 4.3.4.1). *L. aporus* showed the highest number of significant DE orthologous genes compared to

all other species while *L. danicus* and *L. hargravesii* shared the fewest differentially expressed transcripts.

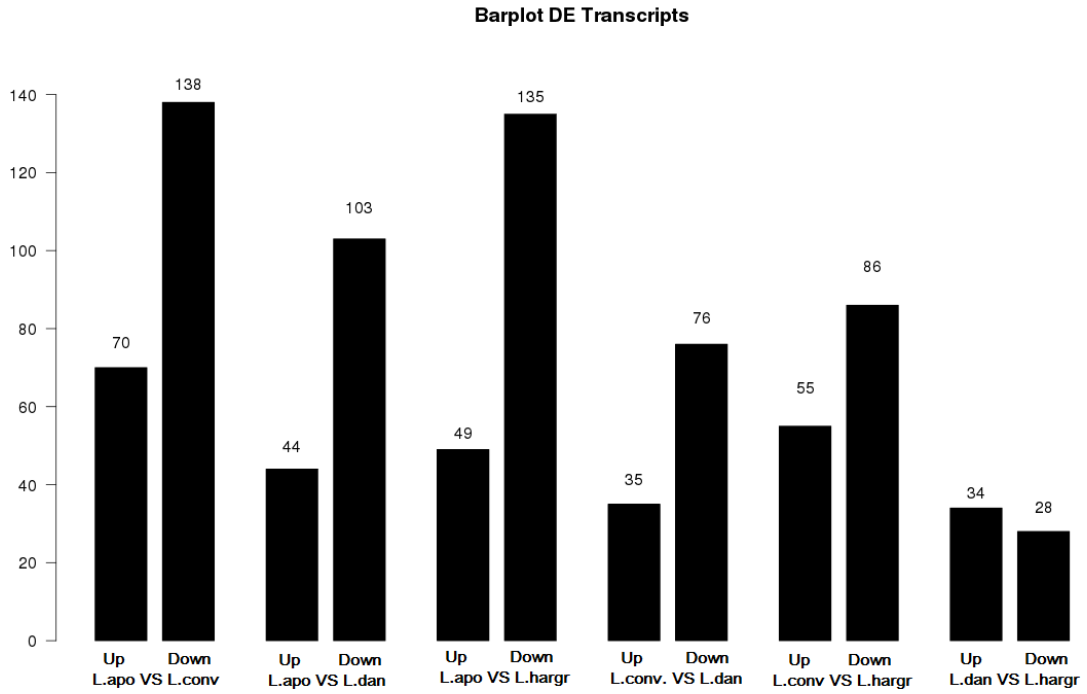


Figure 4.3.4.1 Significantly up and downregulated transcripts (FDR<0.05) of the genes found orthologous across all four species.

In addition, the logarithms of the fold change between the species pairs are presented as boxplots (Fig. 4.3.4.2).

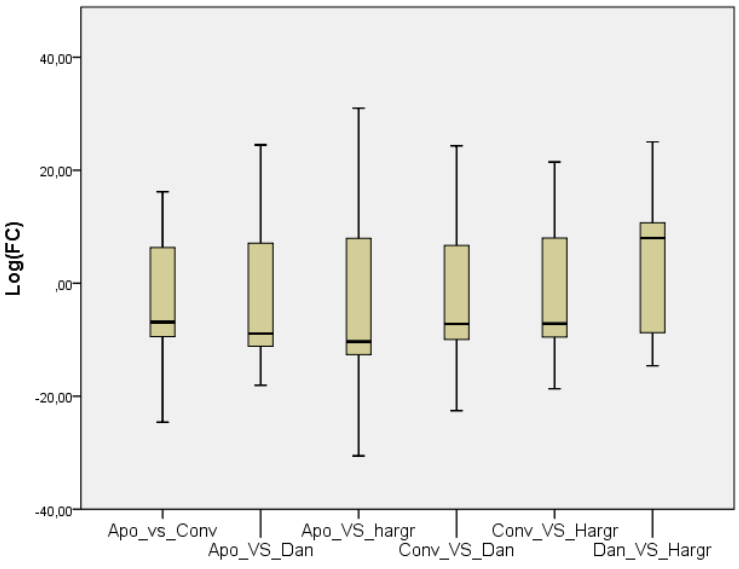


Figure 4.3.4.2 Boxplot of log2(FC) values of the significant DE transcripts detected in the orthologous gene set across all species.

Based on the distribution of the fold change values in the boxplot, the arbitrary values of +15 and -15 were set as upper and lower thresholds to define the highly differentially expressed genes. For

each pair of species, the number of unique significant DE transcripts was also calculated (Table 4.3.4.1 and Fig. 4.3.4.3). The highest differences in expression of orthologous genes were recorded in the *L. aporus* - *L. danicus* pair and the *L. aporus* - *L. hargravesii* pair. Accordingly the highest number of transcripts that were found differentially expressed only in specific pairs belonged to the pairs that involved *L. aporus.*

Table 4.3.4.1 High fold and unique significant DE transcripts of the orthologous genes found across all species.

	L.apo VS *L.conv*		*L.apo* VS *L.dan*		*L.apo* VS *L.hargr*		*L.conv* VS *L.dan*		*L.conv* VS *L.hargr*		*L.dan* VS *L.hargr*	
	High Fold	Unique	High Fold	Unique	High Fold	Unique	High Fold	Unique	High Fold	Unique	High Fold	Unique
Up	2	15	4	7	8	10	2	7	7	18	5	12
Down	7	58	17	28	30	49	9	13	3	27	0	12
Total	9	73	21	35	38	59	11	20	10	45	5	24

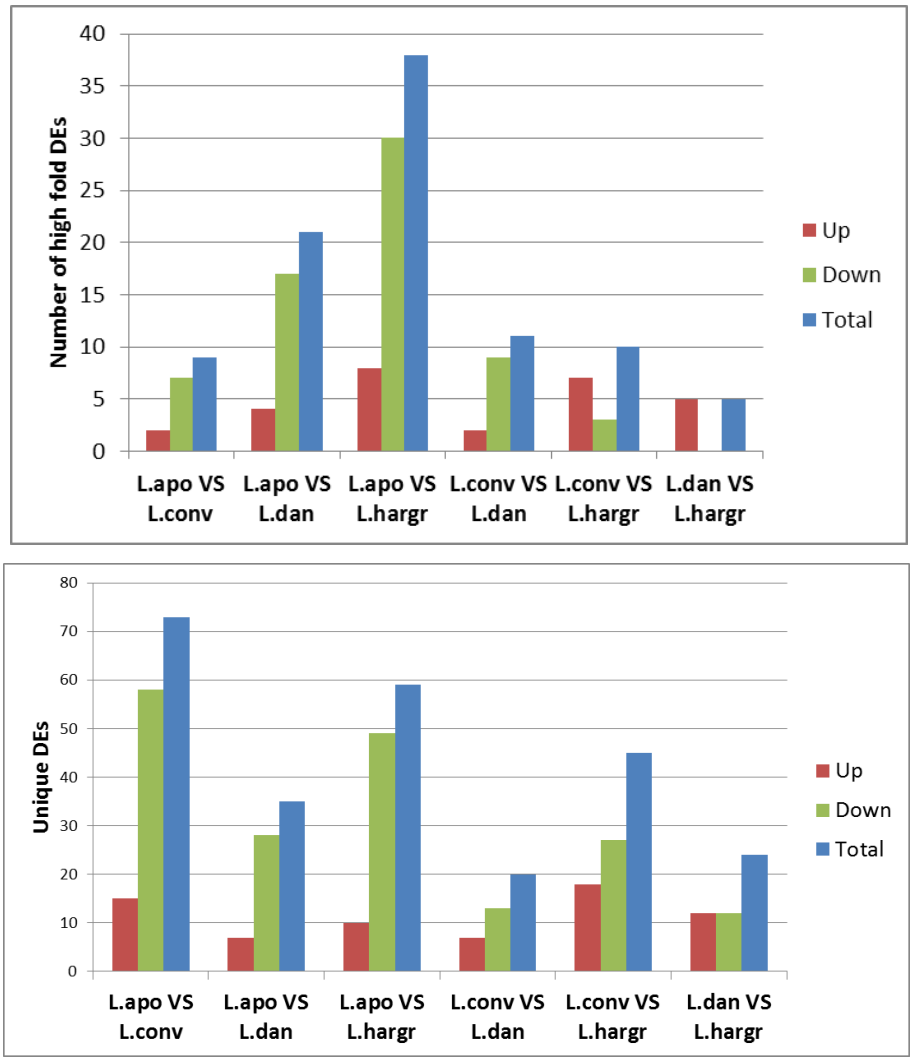


Figure 4.3.4.3. Barplot of high fold (above) and unique (below) significant DE transcripts of orthologous genes found across all species.

The significantly enriched GO terms of the significant DE genes between species indicated the functions for which each pair of species mainly differed. *L. aporus* and *L. danicus* seem to differ in

transportation of molecules, lipid oxidation and regulation of protein catabolism, translation and post translation modifications (Fig. 4.3.4.4).

Figure 4.3.4.4 *L. aporus* vs *L. danicus* biological process GO enrichment (FDR <=0.05) ''TreeMap'' view of REVIGO. Size of the rectangles is adjusted to reflect the GOEA FDR corrected p-value.

Compared to the rest of the species *L. aporus* differed in translation related processes mainly (mRNA polyadenylation, mRNA cleavage) (Fig. 4.3.4.5 -6).

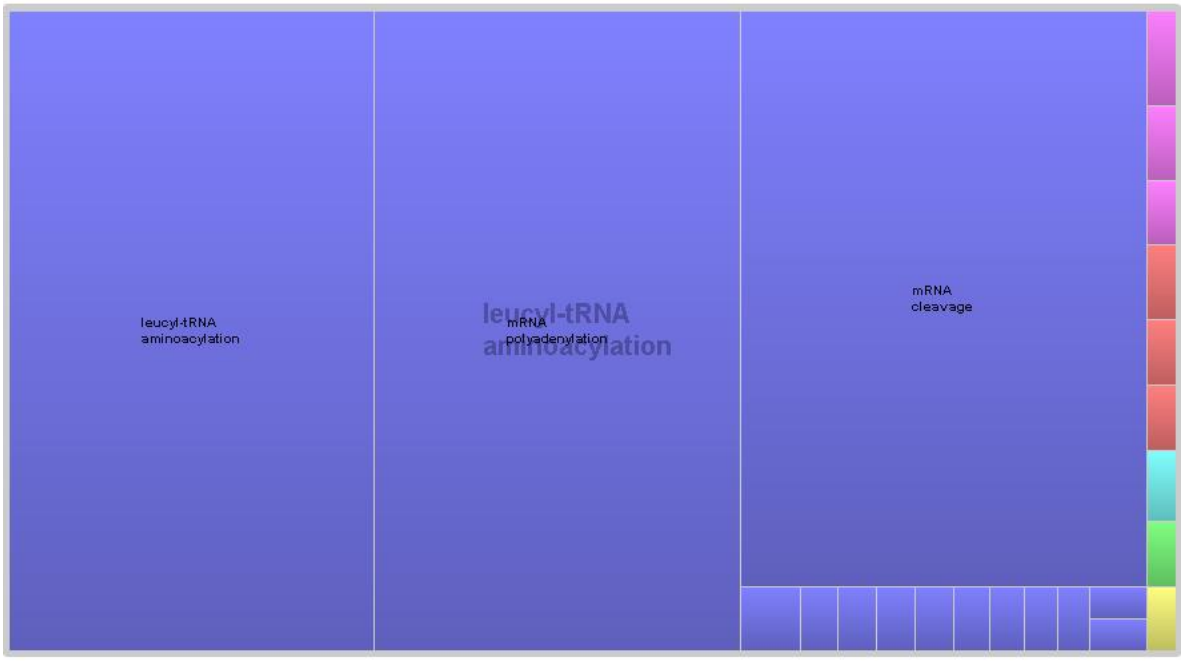


Figure 4.3.4.5 *L. aporus* vs *L. convexus* biological process GO enrichment (FDR <=0.05) ''TreeMap'' view of REVIGO. Size of the rectangles is adjusted to reflect the GOEA FDR corrected p-value.

The small purple rectagles in Fig. 4.3.4.6 were related to transport (nucleocytoplasmic, sulfate, vesicle docking, chloride and transmembrane drug transport).

Figure 4.3.4.6 *L. aporus* vs *L. hargravesii* biological process GO enrichment (FDR <=0.05) ''TreeMap'' view of REVIGO. Size of the rectangles is adjusted to reflect the GOEA FDR corrected p-value.

L. convexus and *L. danicus* were mainly different in amino acid synthesis and metabolism, transportation and signal transduction, post translation modification (proteolysis, protein phosphorylation), nitrogen and carbohydrate metabolism and chromatin remodeling (Fig.4.3.4.7).

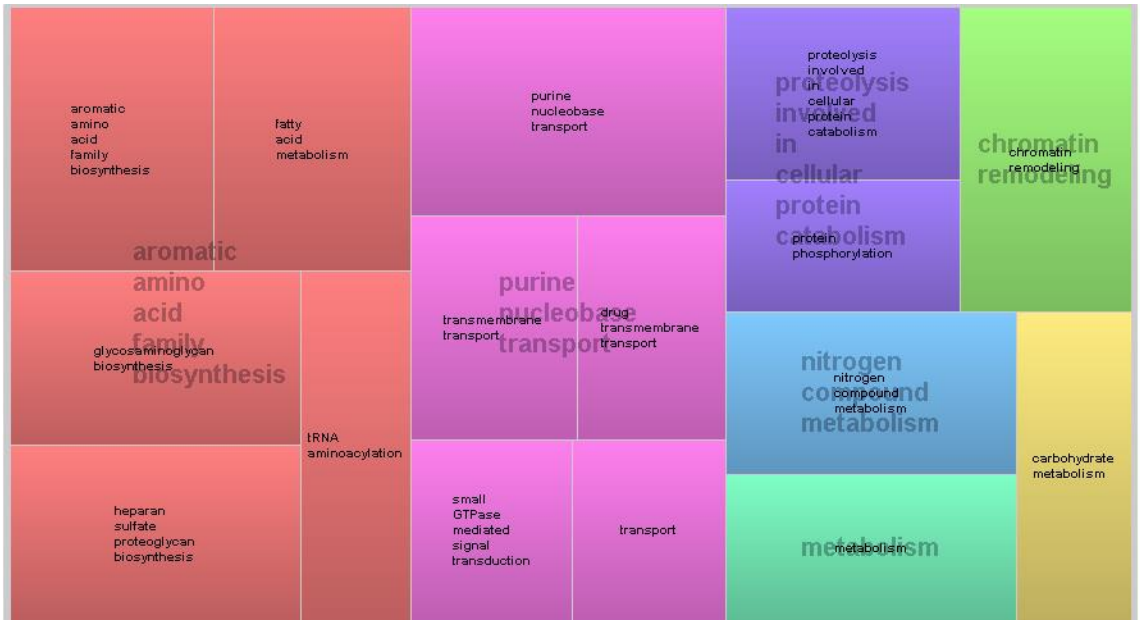


Figure 4.3.4.7 *L. convexus* vs *L. danicus* biological process GO enrichment (FDR <=0.05) ''TreeMap'' view of REVIGO. Size of the rectangles is adjusted to reflect the GOEA FDR corrected p-value.

L. convexus and *L. hargravesii* differed in transportation of proteins and molecules, signal transduction, protein synthesis and metabolism, then in autophagy (degradation and recycling of unnecessary or dysfunctional cellular components), formation and metabolism of sugars (Fig.4.3.4.8).

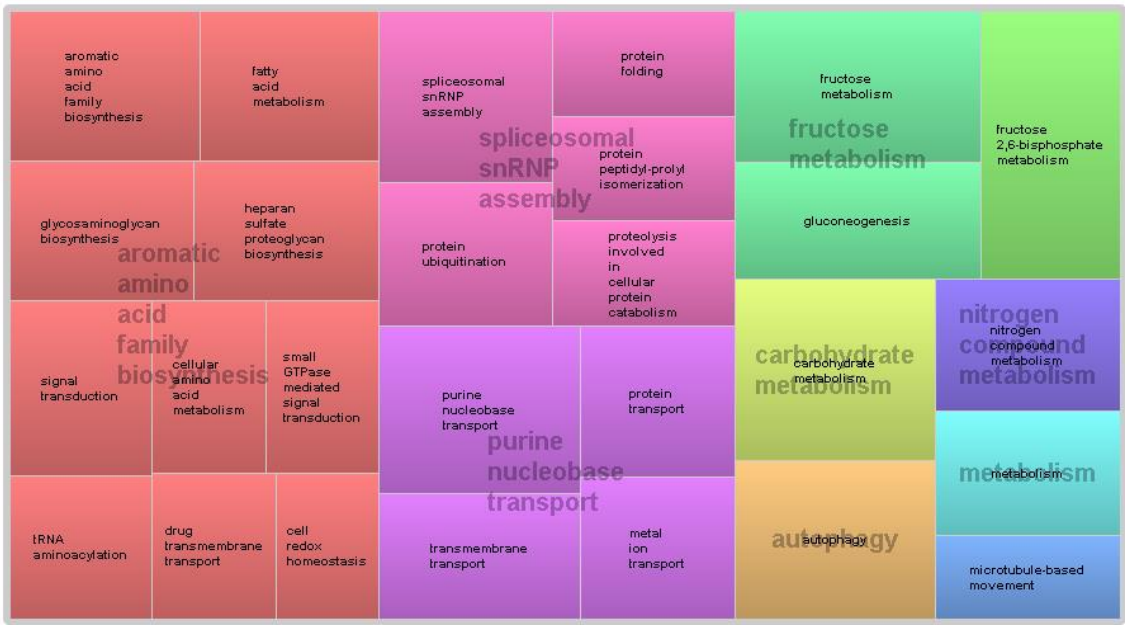


Figure 4.3.4.8 *L. convexus* vs *L. hargravesii* biological process GO enrichment (FDR <=0.05) ''TreeMap'' view of REVIGO. Size of the rectangles is adjusted to reflect the GOEA FDR corrected p-value.

L. danicus and *L. hargravesii* differed mainly in the metabolism of coenzyme A which is notable for its role in the synthesis and oxidation of fatty acids and the oxidation of pyruvate in the citric acid cycle (Fig.4.3.4.9).

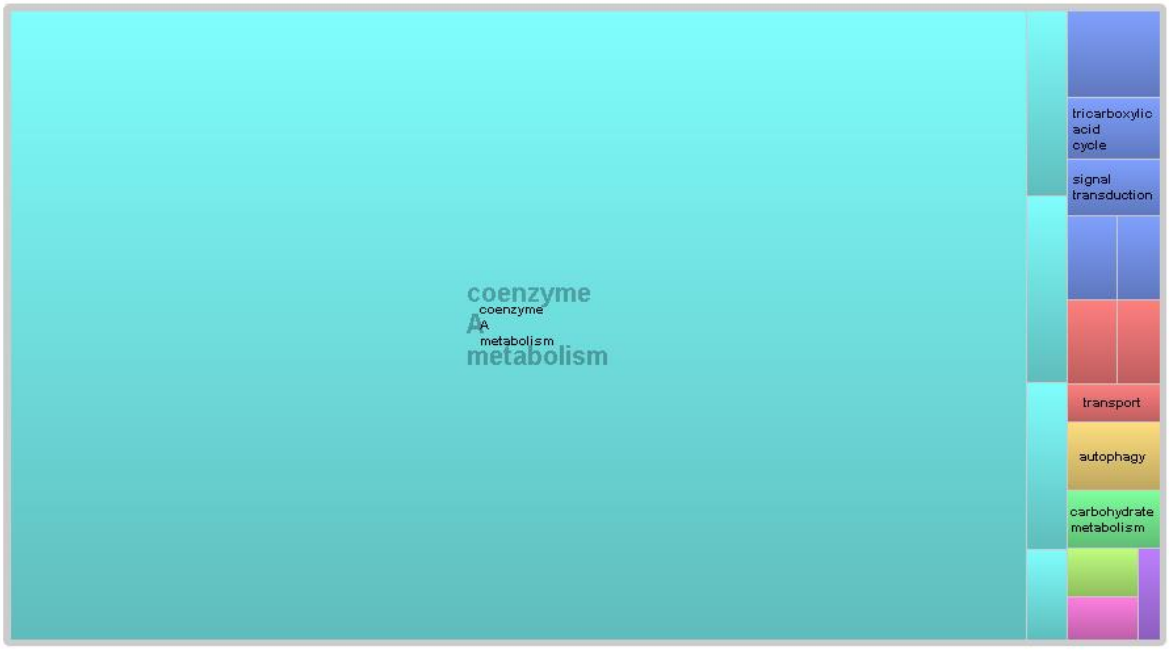


Figure 4.3.4.9 *L. danicus* vs *L. hargravesii* biological process GO enrichment (FDR <=0.05) ''TreeMap'' view of REVIGO. Size of the rectangles is adjusted to reflect the GOEA FDR corrected p-value.

4.3.5. Search for Genes of Interest and TE analysis

The transcripts, which were listed in Materials and Methods, were selected to be searched for in all *Leptocylindrus* transcriptomes because they were already found significantly differentially expressed in the DE analysis of *L. aporus* in the three different temperatures (Chapter 3). The Chapter 3 results makes these genes interesting targets for investigating species functional diversity related to response and adaptation to different environmental conditions. However, the blast search returned no significant results, so no homologous transcripts of *L. aporus* were found in the other species. The flagellar genes that were previously found in *L. danicus* by Nanjappa et al. (submitted) were also detected here in *L. danicus* and *L. hargravesii,* the two species that have been observed to undergo sexual reproduction. The flagellar genes were expected to be expressed in *L. danicus* transcriptome since at least one of the strains, 1089-17, was observed to form flagellated sperms during sexual reproduction in cultures used to obtain RNA for transcriptomics. Sperms and spores were also observed in the other strains but in a much lower degree. The same cannot be said with certainty for any of the strains of *L. hargravesii* but the opposite cannot be excluded either; sperms could have been present and simply missed during the microscopy observation.

The annotation of all transcripts was also screened for transposon-related terms and significant differences can be spotted among the species. The results are shown below:

Table 4.3.5.1 Search hits of transposon-related terms in the transcriptome annotation of each species.

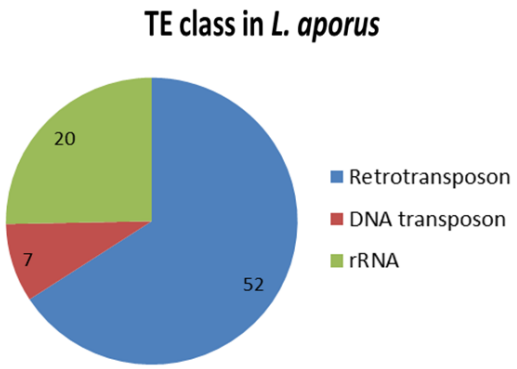
Search Term	Number of Hits			
	L. aporus	*L. danicus*	*L. hargravesii*	*L. convexus*
Retrotransposon	0	0	0	2
Transposon	0	2	2	2
RNase	32	19	17	18
Copia	1	0	0	0
Reverse Transcriptase	151	57	61	27
Transposase	35	23	19	15
Integrase	38	28	16	10
PiggyBac	11	0	0	5
Total	268	129	115	79

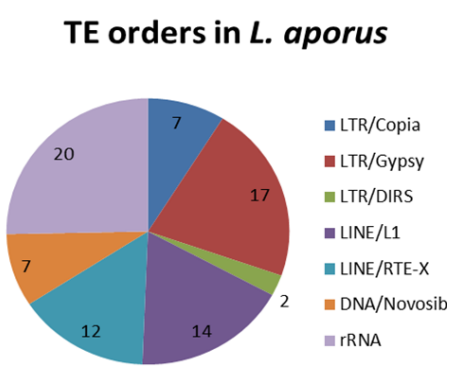
The temperature related transcripts found in the annotation were all annotated as "low temperature requirement" or "low temperature viability protein" (Table 4.3.5.2).

Table 4.3.5.2 Search hits of temperature related terms in each species transcriptome annotation.

Search Term	Number of Hits			
	L. aporus	*L. danicus*	*L. hargravesii*	*L. convexus*
Temperature	12	4	4	6
Heat Shock	436	371	217	251
HSF	348	284	148	174
HSP	79	77	72	52
Cold-Shock	8	16	24	8
Chaperon	108	66	61	62

The specific TE analysis annotated 79 transcripts as transposon related in *L. aporus*, 59 in *L. hargravesii*, 37 in *L. convexus* and 93 in *L. danicus* (Fig. 4.3.5.1).





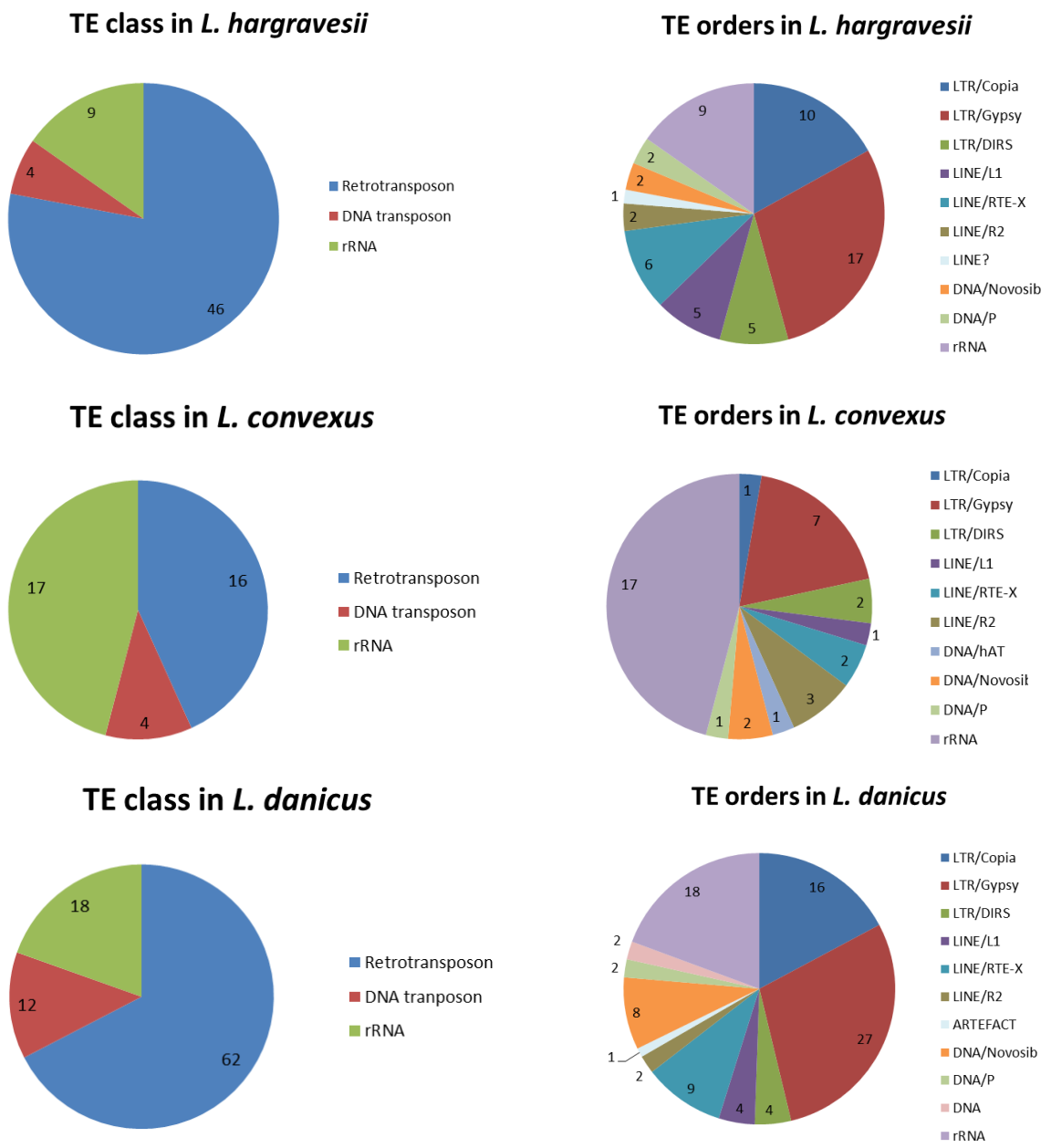


Figure 4.3.5.1. Distribution of transposon related transcripts on class (left) and order (right) level in all four *Leptocylindrus* species.

In all species retrotransposons dominated the TE related transcripts. The species with the lowest number of retrotransposons was *L. convexus* while *L. danicus* and *L. aporus* showed the highest ones. In addition, transposons differentially expressed among strains were mainly detected in *L. hargravesii* and *L. danicus* (Fig. 4.3.5.2).

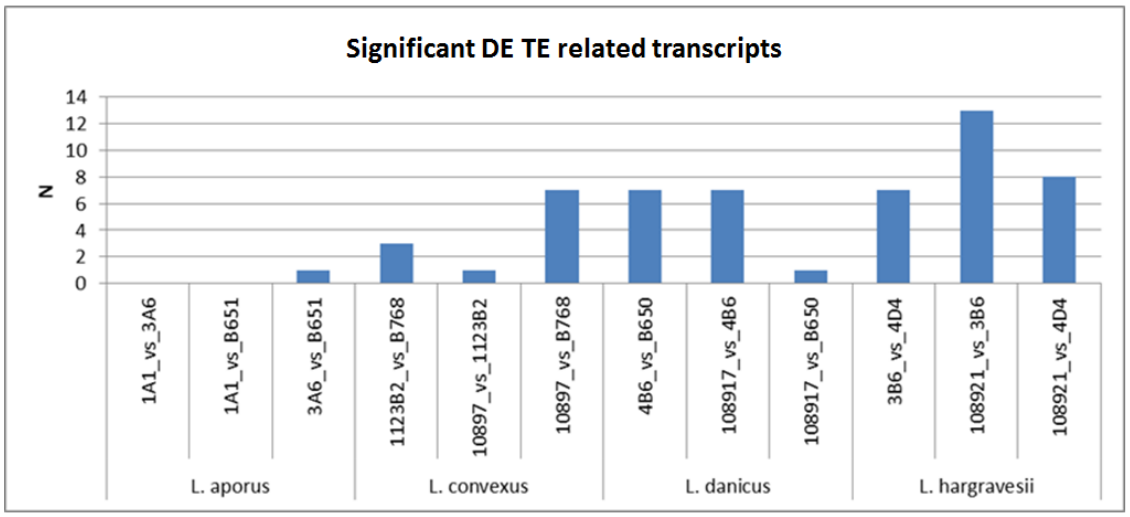


Figure 4.3.5.2 Transcripts related to transposable elements found significantly differentially expressed among strains in each *Leptocylindrus* species.

The variants linked to the TE related transcripts are shown in Table 4.3.5.3. The majority of the variants were SNPs. The species with the higher number of variants was *L. hargravesii*, matching the trend found in the significant DE transcripts.

Table 4.3.5.3 Variants linked to the TE related transcripts found in each *Leptocylindrus* species.

	L. aporus	***L. convexus***	***L. danicus***	***L. hargravesii***
SNPs	143	215	325	371
indels	8	11	3	76

4.4. Discussion

The aim of this part of the thesis was to obtain the transcriptomes of the four *Leptocylindrus* species present in GoN in order to investigate their intra-specific and inter-specific differences through a comparative approach conducted at both expression and sequence level. In addition, the origin and evolution of polymorphisms as remarkable divergence in the four species' traits (phylogenomic study) was explored and took advantage of a set of different analyses such as identification, characterization, and quantification of sequence similarity. Overall, all the analyses performed produced strong evidence for a notable functional diversity within a genus otherwise characterized by a high morphological homogeneity. Differences were found at the level of genomic micro-variations, which could have important consequences for species plasticity and adaptation, as well as at the level of the expression of orthologous genes. As detailed in the following discussion, a major outcome of the analyses was the finding of remarkable diversity and variability even among strains of the same species, which so far has not been taken into much

consideration in transcriptomic research, except in few studies (Llinás et al., 2006; Dugar et al., 2013).

4.4.1. General characteristics of *Leptocylindrus* species transcriptomes

In contrast to genome size, transcriptome size can vary greatly not only among different species but also among and within cell types of the same organism depending on cell size, stage of the cell cycle, ploidy level, age, stress state and growth condition (Coate and Doyle, 2015). Therefore, transcriptome size cannot be considered a direct reflection of the size of the genome but rather of the activity of the cells at a specific time under the given conditions. The following table shows the transcriptome sizes of some representative diatom species, including species under strong perturbations.

Table 4.4.1.1 Transcriptome size of representative diatom species including species under strong perturbations.

Species	Transcriptome size (N of transcripts)	Reference
Pseudo-nitzschia arenysensis	19,852	Di Dato et al., 2015
Pseudo-nitzschia delicatissima	17,595	Di Dato et al., 2015
Pseudo-nitzschia multistriata	21,632	Di Dato et al., 2015
Pseudo-nitzschia tricornutum under nitrogen stress	10,234	Levitan et al., 2015
Thalassiosira pseudonana under nitrate limitation	11,390	Bender et al., 2014
Fragilariopsis cylindrus under nitrate limitation starvation	18,077	Bender et al., 2014
Pseudo-nitzschiamultiseries under nitrate limitation	19,703	Bender et al., 2014
Nitzschia sp. under different salinity conditions limitation	19,430	Cheng et al., 2014

Comparing the *Leptocylindrus* transcriptome sizes to these of other diatoms it can be concluded that the numbers are very close especially for *L. convexus* (18,878 transcripts) and *L. hargravesii* (24,364). The other two species, *L. danicus* (31,806 transcripts) and *L. aporus* (33,434 transcripts), showed a higher number of transcripts than the one expected in diatoms. The many different strains used for the obtainment of the transcriptomes and the high intraspecific variability of these specific species might have influenced this result. Within *Leptocylindrus,* the different size of each species transcriptome was also an interesting result considering that *L. danicus* and *L. hargravesii* are very closely related -morphologically and phylogenetically- and they could have been expected to be the ones with the more similar transcriptome size. But as stated already

transcriptome size depends on many factors. Instead, *L. danicus* and *L. aporus* were the two species with the widest seasonal distribution so it would make sense to be also the ones selected for a larger functional repertoire. Although this is an indication for each species separate evolution and functional capacity, it should not be forgotten that many reads were removed during the quality control and the numbers differed among species based on the RNA and sequencing quality.

4.4.2. Intra- and interspecific genetic diversity based on micro-variations

The variant calling analysis led to the identification of nucleotide differences among the individual strains and the assembled transcriptome for each species, which gave the opportunity to explore partially their genetic diversity with no need of sequencing any part of their genome. The detected variants were grouped to single nucleotide polymorphisms (SNPs), insertions and deletions.

The level of genetic variability was high but variable among the species and within each of them. A comparable level of genetic diversity was found among strains of *L. convexus* and among strains of *L. hargravesii* whereas in the other two species some strains were more variable, namely, *L. danicus* 4B6 and *L. aporus* B651, than the others and quite different from them. For all species insertions or deletions (indels) were a minimal part (0.5 – 1%) of total variants, which were instead dominated by SNPs (35 – 60%). Indel polymorphism is known to have a stronger impact on phenotype because this kind of mutation is more harmful than SNPs (Hamblin and Di Rienzo, 2000). It is also known that indels mainly occur in the loop regions of the affected proteins and are more strongly related with functional changes than SNPs. For that reason they are the predominant evolutionary factor when it comes to structural changes in proteins. Indels are more common in essential proteins, highly connected in protein interaction networks and are especially likely to be involved in intermolecular interactions and species-specific adaptations (Chan et al., 2007; Romero et al., 2006; Hormozdiari et al., 2009; Kim and Guo, 2010). The structural changes lead to the modification of protein interaction interfaces rewiring the interaction networks (Hormozdiari et al., 2009). Therefore indels are under a stronger selection pressure which makes

them a less frequently detected polymorphism than SNPs. The impact of SNPs is low or moderate or even null when the polymorphism does not result in a different amino acid in the protein sequence (synonymous variant). However, SNPs can also be of high impact; they can influence gene expression via effects on DNA methylation, which is an important mechanism of epigenetics (Bell et al., 2011; Gutierrez-Arcelus et al., 2013). Cytosines in CpG sites of the genome (regions of DNA where a cytosine nucleotide is followed by a guanine nucleotide in the linear sequence of bases along its 5' → 3' direction) can be methylated but these sites can be altered by SNPs leading to a transition of cytosine to thymidine, ultimately blocking the methylation (Shoemaker et al., 2010). The effect can also be indirect by altering transcription factor binding, which then independently affects gene expression and DNA methylation levels. In all species there was a high number of synonymous but also missense variants (point mutation leading to a different aminoacid); yet they were all characterized as moderate impact variants.

The variants with the highest impact, and probably mainly associated to indels, were equally abundant in *L. danicus*, *L. convexus* and *L. hargravesii* (if 4B6 is excluded) and less in *L. aporus*. On the other hand, moderate impact variants, which were mainly missense variants and therefore attributed to SNPs, showed a different pattern for *L. hargravesii*, with almost as low SNPs as *L. aporus.* Indeed *L. hargravesii* was the species with the highest percentage of indels and high impact variants while *L. aporus* and *L. danicus* had the less. The different composition percentages of variants in *L. hargravesii* compared to *L. danicus* is an evidence of the possible mechanism that actually differentiates these two cryptic species.

As for intraspecific variability, it is noteworthy that both *L. danicus* and *L. aporus* showed remarkable differences in one of the three strains analysed. A higher number of strains should be analysed here to draw some general conclusion, yet it is tempting to hypothesize that this higher intraspecific variability of *L. aporus* and *L. danicus* could increase the species plasticity leading to broader seasonal distribution compared to *L. convexus* and *L. hargravesii* or that the variability could reflect the larger population size of the two former species.

In all species except *L. hargravesii* there is one strain which comes from the culture collection SZN (B650 in *L. danicus,* B651 in *L. aporus,* B768 in *L. convexus*). There is no clear pattern though differentiating these strains within each species, except for the *L. aporus* old strain, which means that either the *L. aporus* strain was the only one affected by the in-culture evolution or that the differences seen in this analysis are mainly based on the properties of the strains rather on the effect of their maintenance conditions. Indeed, the *L. danicus* strain 4B6, which mostly differentiated from the rest *L. danicus* was a recently isolated strain.

In order to gain a more complete idea of the impact of these micro-variations on each species performance, the functions related to the transcripts affected by the high impact variants, so by indels, were investigated. The annotation result showing more unique functions affected by variants for *L. danicus* and *L. convexus* could have been influenced by the much higher percentage of annotated high variants in these species compared to the other two. The incomplete annotation of transcriptome is indeed a problem to be taken into consideration, as a full annotation might have given a completely different pattern. The biological processes related to the shared high impact variants of *L. danicus* and *L. convexus* might be a sign of higher flexibility in the mobility of signal proteins and therefore higher interaction possibilities of these species with external or internal stimuli, such as stress. Nevertheless, most of the functions affected by the variants were apparently the same in all species, supporting what was mentioned previously about indels having an impact on essential proteins that are involved in intermolecular interactions and ultimately species-specific adaptations (Chan et al., 2007; Hormozdiari et al., 2009). Transportation of molecules, signals and proteins seems to be the main target of variants in *Leptocylindrus* species so probably the key pathways for the plasticity of the genus lie among them.

In addition to the exploration of the variants diversity and the related functions affected, a phylogenetic tree based on the sequences of the orthologous (shared between species) genes among species offered another opportunity to investigate the genetic diversity of the species. The tree showed that *L. danicus* and *L. hargravesii* were the most closely related species while *L.*

danicus and *L. aporus* had the highest intraspecific variability. Yet the tree was different from the phylogenetic tree produced based on the ITS marker (the same phylogenetic relationships have been inferred by six more genetic markers, Nanjappa et al., 2013) regarding the grouping of *L. aporus* and *L. convexus*. This mismatch between the two trees could be a result of the alignment-free calculation used in the orthologous based tree. Although the method used here is significantly more accurate than other available alignment-free methods, there might still be a decrease in phylogenetic accuracy (Höhl and Ragan, 2007). With the increase of genomic and transcriptomic data produced, the emphasis of phylogenetic inference has shifted from the search for informative characters to the development of better reconstruction methods for using all these data. Phylogenomic reconstruction methods can be divided into (a) sequence-based methods where primary sequences are compared and trees are built from multiple-sequence alignments and (b) methods based on whole-genome features. In the first method, the individual genes are concatenated and a 'supermatrix' is created, based on which a tree is built (Felsenstein, 2004). Alternatively, each gene can be analysed separately and the resulting trees can be combined to a final 'supertree' (Bininda-Emonds, et al. 2002). In the second method, no alignment is required since it takes advantage of the gene content, gene order or the distribution of oligonucleotides ('DNA strings') (Lin and Gerstein, 2000; Pride et al., 2003). Free-alignment methods are generally considered as a better choice since genes and genomes are subject to recombination, rearrangement and lateral genetic transfer and therefore homologous positions cannot be assumed to always occur in the same order relative to one another (Wong et al., 2008; Wu et al., 2012). In the present analysis, the tree was built with free-alignment method, in particular a pattern-based approach (Höhl et al., 2006). In this case, pattern discovery was used to find regions of similarity occurring in two or more sequences with no alignment necessity. Sub-sequences with shared properties such as identity or match length are extracted and used to compute a distance matrix (Chan et al., 2014). The similarity of the subsequences (profiles) is converted into measures representing the evolutionary relatedness between two full-length sequences and phylogenetic relationships can be computed. These steps are novel in the

alignment-free context and have been proven to significantly improve the overall reconstruction accuracy (Höhl and Ragan, 2007). In particular, it was suggested that alignment free methods are more robust compared to multiple sequence alignment methods especially against among-site rate heterogeneity, compositional biases, genetic rearrangements and insertions/deletions but more sensitive to recent sequence divergence and the presence of incomplete (partial or fragmented) sequence data (Chan et al., 2014). Therefore, the reliability of this method really depends on the nature of the data. In addition to the method itself, it should not be forgotten that the orthologous tree was a result of a concatenation of many different genes. As mentioned in the Introduction, each gene might follow a different evolutionary path depending on the selection pressure that is applied to each one of them. Genes that are part of standard functions are more conserved and thus more reliable for the construction of a true species tree while genes involved in functions such as metabolic processes in response to the environment are more variable and probably more suitable for intraspecific trees. Indeed, in a preliminary exploration of specific genes, phylogenetic trees of four out of five ribosomal proteins found orthologous among *Leptocylindrus* species agreed with the ITS based tree. Therefore, the tree produced by all the genes has a potential noisy background regarding the species phylogeny reconstruction.

4.4.3. Intraspecific functional diversity

After exploring the diversity based on genetic micro-variations, the gene expression patterns were also investigated with the aim to understand if functional diversity reflects the detected genetic diversity. Indeed, the pattern seen in the variant calling analysis was also met in the DE expression analysis with *L. aporus* showing the lowest intraspecific DE level. The higher difference observed in the pair 1089-17 - 4B6 in *L. danicus* could be ascribed to 4B6 (4B6 –B650 pair shows also slightly more DEs), which had already been found divergent in the variant calling analysis. The difference cannot be related to the different isolation or filtration periods since in terms of isolation 4B6 and 1089-17 were both winter strains and B650 was a summer strain, and in terms of filtration period 1089-17 was filtered in winter while the other two in late spring. So if any of them were different due to seasonality then it should have been either B650 or 1089-17. But this was not the case.

Similarly the 1A1-3A6 pair in *L. aporus* had more DEs than the other *L. aporus* pairs implying that 1A1 and 3A6 were equally different from B651 but more diverse to each other. This is an interesting outcome when compared to the result of the DE analysis in relation to temperature variations (Chapter 3), which was done with different bioinformatics tools. In temperature experiments, the expression patterns of 1A1 and 3A6 were again equally different from B651, but the difference between the two of them was the lowest. This contrasting result highlights how different methods and tools can lead to different results and interpretations (Seyednasrollah et al., 2013). The technology and tools for transcriptome analysis continue to evolve and so far the agreement between results obtained from different tools is still unsatisfactory; results are affected by parameter settings, especially for genes expressed at low level (Conesa et al., 2016). However, certain methods might work better for certain datasets based on specific criteria. To this end, users should get familiar with the general relationships between and within sample groups using quality assessment and general visualization methods before selecting the analysis tool. For example, in the case of the comparative transcriptomics analysis performed in this chapter, NOIseq was not the best tool for all species, but EBseq (a more conservative package for DE analysis) could have been used for *L. aporus* and *L. danicus*. So eventually, the need to use a tool that could be applied to all species, and thus produce comparable results, dictated the choice of NOIseq. Of course the differences with Chapter 3 may lie in the other steps of the analysis as well, e.g. the assembly and the annotation step. Indeed, the difference in annotation achieved in the two different analyses was notable. This could be a result of a better assembly of the one over the other leading to more coding sequences that were able to be annotated. But this possibility is low since in both assemblies Trinity was one of the main tools and therefore their quality should be quite similar. However, during functional annotation the procedures differed more. In the current chapter, the high quality transcripts were translated into proteins with *Transdecoder* tool and then the coding sequences obtained were functionally annotated with the InterPro database. In addition to InterPro annotations, using protein sequences, a BlastP analysis was performed against NCBI in order to functionally annotate the proteins by sequence similarity. On the other

hand, Annocript that was used in chapter 3 utilizes a different range of databases for annotation including Swiss-Prot, UniRef90, the Conserved Domains Database (CDD), Rfam and SILVA database. The latter database, which is the one completely missing from the annotation analysis of this chapter, might have contributed to this discrepancy between the analyses. The use of Annocript for annotating the assembly produced in this chapter as well could answer this question. Lastly, the fact that samples of the strains grown at low and high temperature were used as replicates in Chapter 3, possibly increased the possibility to detect larger differences in the gene expression levels among strains. Nevertheless, the DE analysis of this chapter revealed that there were significant differences between the different strains of *L. danicus* and *L. aporus*, as concluded also by the variant calling analysis.

What was also quite informative was the low number of significant DE transcripts shared between species pairs which could support the notion that different components of the pathways were differently regulated between the strains, due to either mutations in regulation regions or epigenetic variations. The pattern seen in the high fold DE transcripts followed the high impact and indel variant pattern described before, with *L. hargravesii* and *L. aporus* holding the highest and lowest percentage respectively. A link of the diversity offered by the variants and the functional patterns could be assumed at this point.

The analysis of the annotated differentially expressed transcripts could shed more light on these details but before proceeding with it, it should be reminded again that the level of unannotated DE transcripts between species and even between strains varied, especially within *L. danicus* with the DE transcripts of 4B6 – B650 including up to almost 90% of unannotated transcripts, whereas within *L. aporus* with 3A6 – B651 pair missing annotation for 60% of significant DE transcripts. In a study on the transcriptional response of three diatoms – *Thalassiosira pseudonana, Fragilariopsis cylindrus and Pseudo-nitzschia multiseries* - the majority of the genes found significantly differentially expressed had minimal or no annotation information (Bender et al., 2014), while Di Dato et al. (2015) noticed that in *Pseudo-nitzschia* the species-specific protein list contained mainly proteins associated to unknown functions together with proteins associated to known

functions, present in all species. These authors suggested that the unique proteins associated to the common functions are divergent proteins belonging to related gene families and are likely responsible for the plasticity of the species; this might also be the case here for the unannotated proteins, especially considering the agreement between the variant calling and DE analysis results.

Nonetheless, the annotated transcripts offered valuable information for each species. In *L. aporus,* DNA integration related transcripts were not highly enriched as in the results of transcriptomic responses under different temperature conditions (Chapter 3) but they were still present. In *L. danicus*, as in *L. aporus,* terms related to transportation were enriched but post transcription regulation terms were also present in pairs where B650 was involved. This could be a strain specific difference again, maybe related to the summer origin of B650 or it could be related to the fact that B650 is the old strain coming from the SZN culture collection and in-culture evolution could have led to the modification of its regulatory mechanisms in order to adjust its functions to the new stable environment. In the 1089-17 – 4B6 couple, lipid biosynthesis was highly enriched. Lipid metabolism is related to membrane fluidity so these strains could differ in the maintenance of the general membrane state under environmental challenges or again in the signal transduction process. The cell division functions mostly enriched in the *L. convexus* pairs including 1123B2 showed that this strain was probably more active in cell division compared to the other strains, which could be related to the fact that it was the most recently isolated or that it was a summer strain. Following this reasoning, 1123B2 could be still more active in replicating compared to the slow growing acclimatized/winter strains. In all three *L. hargravesii* pairs the biggest differences were related to quite basic biological processes that nevertheless were diversified between the strains. Lipid biosynthesis was highly enriched in the differentially expressed transcripts between 1089-21 and 3B6 *L. hargravesii* strains implying differences in membrane fluidity or signaling. Overall, functional differences could be detected among the strains of all species with some of them differentiating more than others in specific terms.

The gene expression results are summarized and presented along with similar results of other transcriptomic studies on diatoms in table 4.4.3.1.

Table 4.4.3.1 Summary of statistics on the genes found significantly differentially expressed among strains (only the unique DE transcript have been indicated, Table 4.3.2.1) and species in *Leptocylindrus*, as well as in perturbation studies of other diatom species.

Species / strains	N of DE genes		log2FC range	Significance threshold	% to total transcripts		Reference
	Up	Down			Up	Down	
Leptocylindrus aporus , 1A1 vs B651	28	80	(-8) - (7)	FDR < 0,05	0,08	0,24	Current study
Leptocylindrus aporus , 3A6 vs B651	101	94		FDR < 0,05	0,30	0,28	
Leptocylindrus aporus , 1A1 vs 3A6	309	238		FDR < 0,05	0,92	0,71	
Leptocylindrus danicus , 4B6 vs B650	425	516	(-10) - (10)	FDR < 0,05	1,34	1,62	
Leptocylindrus danicus , 1089-17 vs B650	582	229		FDR < 0,05	1,83	0,72	
Leptocylindrus danicus , 1089-17 VS 4B6	1150	633		FDR < 0,05	3,62	1,99	
Leptocylindrus hargravesii , 1089-21 vs 3B6	658	566	(-14) - (14)	FDR < 0,05	2,70	2,32	
Leptocylindrus hargravesii , 3B6 vs 4D4	767	540		FDR < 0,05	3,15	2,22	
Leptocylindrus hargravesii , 1089-21 vs 4D4	421	436		FDR < 0,05	1,73	1,79	
Leptocylindrus convexus , 1089-7 vs 1123B2	382	346	(-10) - (10)	FDR < 0,05	2,02	1,83	
Leptocylindrus convexus , 1089-7 vs B768	505	548		FDR < 0,05	2,68	2,90	
Leptocylindrus convexus , 1123B2 vs B768	312	455		FDR < 0,05	1,65	2,41	
L. aporus vs *L. convexus* *	70	138	(-17) - (17)	FDR < 0,05	4,33	8,54	
L. aporus vs *L. danicus* *	44	103	(-18) - (12)	FDR < 0,05	2,72	6,37	
L. aporus vs *L. hargravesii* *	49	135	(-26) - (17)	FDR < 0,05	3,03	8,35	
L. convexus vs *L. hargravesii* *	55	86	(-15) - (20)	FDR < 0,05	3,40	5,32	
L. convexus vs *L. danicus* *	35	76	(-18) - (15)	FDR < 0,05	2,17	4,70	
L. danicus vs *L. hargravesii* *	34	28	(-15) - (17)	FDR < 0,05	2,10	1,73	
Phaeodactylum tricornutum under nitrogen stress	2754	2866	(-10) - (10)	log2FC ≥ 2 and FDR < 0,05	26,91	28	Levitan et al., 2015
Thalassiosira pseudonana under nitrate limitation	787	1319	(-8) - (8)	P-value < 0,01	7	11,65	Bender et al., 2014
Fragilariopsis cylindrus under nitrate limitation	285	731		P-value < 0,01	1,64	4,2	
Pseudo-nitzschiamultiseries under nitrate limitation	737	1070		P-value < 0,01	4,38	6,35	
Nitzschia sp. under different salinity conditions	3634	3962	(-30) - 30	log2FC ≥ 2 and FDR < 0,05	18,70	20,39	Cheng et al., 2014
Thalassiosira pseudonana under phosphorus stress	1382		(-3) - (12)	FDR < 0,05	14,44		Dyhrman et al., 2012
Thalassiosira oceanica under iron limitation	300		(-5) - (5)	no p-value, log-likelihood ratio test statistic	2,66		Lommer et al., 2012
Thalassiosira pseudonana under low iron availability	632	680	(-4) - (5)	log2FC ≥ 2 and p-value < 0,05	5,58	6	Thamatrakoln et al., 2012
Thalassiosira pseudonana under silicon limitation	709		(-3) - 3	log2FC ≥ 2 and Bayesian t-test P <	8,89		Mock et al., 2008
Phaeodactylum tricornutum under iron starvation	212	26	(no info) - (4)	log-likelihood ratio test	2,45	0,3	Allen et al., 2008
* DE analysis between genes found orthologous among all *Leptocylindrus* species							

Despite the different techniques and methods used in the studies on other diatoms, the percentage of the genes that respond significantly at different conditions are comparable to the interspecies changes seen between *Leptocylindrus* species but most importantly also to the intraspecific variability of the individual species. Especially in the cases of *L. danicus* and *L.*

hargravesii, the number of genes found significantly differentially expressed among strains is close to or even higher than the number of DE genes detected in diatom species under strong perturbation e.g. in *F. cylindrus* under nitrate limitation, *T. oceanica* and *P. tricornutum* under iron limitation. This is clear also comparing the results of the intraspecific expression analysis with the interspecific one in the current study, where the same techniques and methods were used. Although in general the interspecific changes are of a higher degree, there are still cases of strains where the amount of significant differences in gene expression is comparable to the amount of significantly differentially expressed genes between species.

4.4.4. Interspecific functional diversity

In addition to the intraspecific functional diversity, the relationship between species was also explored based on the expression values of the orthologous genes across all species. A general first idea of each species unique, as well as shared, functional traits was formulated. The slight differences observed between the PCA, HCA and CCA of expression values were expected since each method is based on different manipulation of the data, with hierarchical clustering being a classification method based on the dissimilarity of the expression values while PCA (as well as CCA) is an ordination method that tries to map the strains on two or three dimensions in order to reflect the similarity/ dissimilarity of the whole community. CCA was the method that better grouped the species, with almost equal distances between strains. The outcome of this analysis can be far from the known genetic relationship since species can share functions or strategies without being necessarily genetically close. This is indeed confirmed with the outgrouping of the expression levels of *L. convexus* and *L. hargravesii*, which could have some relation with the more restricted seasonality of these species.

Considering the expression variations among species, it was quite reasonable for *L. aporus* to show the highest expression differences as well as the highest number of DE orthologous genes compared to all other species because of its much lower expression level of transcripts. On the other hand, *L. danicus* and *L. hargravesii* shared the fewest differentially expressed transcripts. As a consequence, the pairs involving *L. aporus* had the more unique DE transcripts. In order to

enable a more informative functional interpretation, a gene ontology enrichment analysis was carried out on the differentially expressed transcripts across all the species. The number of functions enriched when comparing *L. aporus* to *L. danicus* was much higher than when comparing *L. aporus* to *L. convexus* or even to *L. hargravesii.* This result was obviously influenced by the high strain variability of *L. aporus* and *L. danicus* compared to the other species, but it still suggests that the wider seasonal distribution of *L. danicus* and *L. aporus* is possibly driven by a wider availability of specific genes or flexibility of the regulatory systems.

In GO term enrichment of all species, it was generally noticed that translation and post translation modifications, as well as signal transduction, and molecule transportation, held a crucial role. This confirms, first of all, that differential regulation of separate components of the same metabolic pathway, which results from post-transcriptional and post-translational modification for selected genes or proteins, respectively, might account for the majority of the differences seen in the species expression patterns. Secondly, the impact of indels in intermolecular interaction discussed above is indeed important in the differentiation among the species. Fatty acid metabolism was also present in the enrichment analysis results between most of the species pairs, proving that the properties of the membrane are an important component of the functional diversification among diatom species, which is probably related to the different environmental conditions each one of them can withstand. The limited number of functions that were enriched in the *L. danicus – L. hargravesii* pair showed that their high genetic similarity was also matched by the expression patterns of their orthologous genes.

At this point, we could sum up the functional strategies of the species and divide them into two levels. The first level controls their wide/year-round or restricted seasonality and the second one regulates the specific environmental preference of each species e.g. summer or winter blooms:

1. The species that are present year-round (*L. aporus* and *L. danicus*) hold a higher within-species diversity compared to the others whereas the low variability between the strains of *L. convexus* and *L. hargravesii* could go hand-in-hand with their more restricted seasonal distribution. This intraspecific diversity is genetic (phylogenetic tree, SNPs and

indels) and likely has an influence on the functional diversity. A recent study by Wohlrab et al. (2016) has demonstrated gene expression changes in a genotype-specific manner in dinoflagellates. As it is stated in that study, "the genotype-specific interactions and associated trait variation within a population maintain disequilibrium among genotypes and can, therefore, among other mechanisms, explain the paradox of plankton".

2. Each species-specific seasonal preference is depicted mainly by the differences in general functions shared by all such as transportation of molecules and signal transduction. This concept is better seen in *L. danicus* and *L. aporus* DE analysis of their orthologous genes.

Additional studies investigating the similarities and differences to different stressors can help determine whether or not the marked diversity in gene-specific responses is linked directly to significantly different ecological consequences in the field (Bender 2014).

4.4.5. Assessment of specific genes

Genes of interest identified in transcriptomics or genetic studies of certain species may turn out to be important and play an equally critical role in other species as well and even be a part of a mechanism highly conserved among many different organisms. Therefore, the transcripts related to heat and DNA integration that were found significantly different under the different temperature conditions in *L. aporus* (Chapter 3) were blasted against the transcriptomes of *L. danicus, L. hargravesii* and *L. convexus*. None of them returned a good match, which could mean that these transcripts are quite species specific or that they are strictly activated by the temperature change.

On the other hand, homologues of flagellar genes previously detected in *L. danicus* (Nanjappa et al., submitted) were also found in *L. hargravesii,* but not in other species confirming that the expression of flagella-related genes is specific to the species known to undergo sexual reproduction, which for the Leptocylindraceae only include *L. danicus* and *L. hargravesii*. However, one might expect that more meiosis-related genes would be expressed in tandem with the flagellar genes, especially for *L. danicus* where at least a strain was highly sexually active. This was not the case when meiosis related terms were searched for in the annotation. In a

comparative analysis of all genes functions for each species separately, information on meiotic genes would possibly be obtained. Indeed, in Nanjappa et al. (submitted), where *L. aporus* and *L. danicus* transcriptomes were annotated and functions were compared by Fisher's exact test, genes associated with the GO terms: cilium, flagellum, cilium morphogenesis, female meiosis, female gonad development, female sex differentiation, development of primary female sexual characteristics, sex differentiation, male meiosis, showed a higher representation whereas in *L. aporus* often the genes associated with these GO terms were absent. In addition, the expression of the genes in *L. hargravesii,* where no clear sperms were observed (though not excluded to be present) when RNA was obtained, could be explained by an early activation of the genes necessary before the actual formation of the sperms or expression of proteins that share homology with the components of the flagellum. The latter case has been proven for about 25 protein components of the bacterial flagellum and the type-three secretion system (TTSS) where one likely evolved from the other or the two structures evolved in parallel (Pallen and Matzke, 2006; Saier, 2004). In addition, several cilia proteins have been found at non-cilia sites, where they have specific functions, indicating the involvement of cilia proteins in diverse, cilia-independent, cellular processes and structures (Vertii et al., 2015). For example, mitotic spindles and primary cilia require the function of microtubule-mediated, motor driven transport for their assembly and, hence, flagellar related proteins, such as intraflagellar transport proteins, could be expressed even when the cell does not undergo sexual reproduction (Serra, 2008; Delaval et al., 2011).

The results of the search for temperature related terms were quite interesting since they also pointed towards a difference among the species. Starting with HSPs and HSFs, HSPs were found in equal number among species while HSFs were much higher in *L. aporus* and *L. danicus* compared to *L. hargravesii* and *L. convexus.* Considering that HSFs are the transcription factors regulating HSPs it might be safe to say that in the case of *Leptocylindrus* species the differences between functions related to HSPs are actually a result of the variation of their regulatory system (the fourth mechanism mentioned in the introduction of the current chapter). As mentioned in

Chapter 3, HSPs are proteins produced in response to exposure to stressful conditions and they contribute to the stabilization of the cell functions under these conditions by the proper folding of proteins. Differences of their regulation factors were found between *Leptocylindrus* species that are year-round (*L. aporus* and *L. danicus*) and species with a more strict seasonality (*L. hargravesii* and *L. convexus*). A more complicated and flexible HSF regulatory system might be one of the essential molecular toolkits for a species in order to cope with the many different stressful environments that they experience during different seasons.

The rest of the terms, which were actually more related to temperature since HSPs and HSFs include other kinds of stress as well, showed a pattern which in this case differentiated *L. aporus* and *L. convexus* from *L. danicus* and *L. hargravesii.* In particular, the low temperature viability proteins (ltv) and low temperature requirement proteins (ltr) were more abundant in *L. aporus,* followed by *L. convexus,* whereas *L. danicus* and *L. hargravesii* showed the lowest number for presence of these terms. Although the names of these proteins might point at some relation with response to low temperature, their exact function is still unknown. Ltv1 is a non-ribosomal protein required for the biogenesis of the 40S ribosome subunit and the pre-ribosomal RNA processing in *Saccharomyces cerevisiae* and *Drosophila*, therefore being required for cell growth by permitting protein synthesis (Loar et al., 2004; Kim et al., 2015). When initially identified, it was proven that at low temperature yeast cells lacking *LTV1* gene grow slowly, are hypersensitive to inhibitors of protein synthesis and have aberrant polyribosome profiles. Seiser et al. (2006) confirmed these results but also proposed that Ltv1 is nonessential and functions as one of the several possible adapter proteins that link the nuclear export machinery to the small ribosome subunit. This protein was actually only present in *L. aporus* and *L. convexus* annotations (2 hits in each one, corresponding to a single transcript) and its expression showed a higher level in *L. convexus* than in *L. aporus*. Although this is a result that could add information to the general expression pattern of the species under investigation, it cannot be concluded that this protein plays an important role in the response of the species to temperature alterations.

The rest of the temperature term hits were low temperature requirement (Ltr) proteins. Not much information is available for these proteins either, except that in *Listeria monocytogenes* they are essential for growth at low temperatures but dispensable at higher temperatures (Zheng and Kathariou, 1994). Five transcripts in *L. aporus* and two transcripts in *L. danicus, L. hargravesii* and *L. convexus* were annotated as Ltr. The highest expressed transcript though was found in *L. hargravesii* while *L. aporus* transcripts showed very low expression levels; *L. convexus* and *L. danicus* had an intermediate expression level. Furthermore, four transcripts were related to cold shock proteins (CSPs) in *L. aporus* and *L. convexus*, eight in *L. danicus* and twelve in *L. hargravesii*. The most highly expressed ones were again found in *L. hargravesii* and the lowest in *L. aporus.* (CSPs) are a highly conserved family of proteins that bind to single-stranded nucleic acids. The bacterial CSPs are mainly induced after a rapid temperature drop and regulate the adaptation to cold stress but also other biological functions under normal conditions (Horn et al., 2007). In plants, some CSPs are also upregulated by exposure to cold, increasing tolerance to freezing (Karlson and Imai, 2003; Kim et al., 2009). However, a recent study revealed that CSPs do not share the same function in dinoflagellates since cold shock does not induce them (Beauchemin et al., 2016). Dinoflagellates though are a particular case since to date, with only one exception, no transcription factor has been described and characterized experimentally and it seems that transcriptional control is not a predominant mechanism for regulating gene expression in this group of protists (Morey et al., 2011; Roy et al., 2014). Assuming that these proteins hold the same role in *Leptocylindrus* as in bacteria and plants, their presence and expression could be linked to the seasonal distribution of cold tolerant species compared to the ones that prefer spring-summer months avoiding low temperatures. The above observations on the species' expression patterns of temperature related proteins offer more evidence that the seasonal preference of each species is reflected in the differential expression of essential proteins.

Finally, even though neither DNA integration- nor any transposon-related function was highly enriched in any of the gene enrichment analyses between strains or species, the search of specific TE terms in the annotation did return different results for each species. *L. aporus* had the highest

number of TE related transcripts, *L. danicus* and *L. hargravesii* followed and *L. convexus* had the lowest number. The in-depth analysis of transposons confirmed the much higher presence of retrotransposons compared to DNA transposons in all species. The Gypsy order (see Appendix 1) was the one dominating in all of them but Copia order followed in *L. hargravesii* and *L. danicus* whereas LINE order was slightly more abundant than Copia in the other two species. In any case *L. convexus* was indeed the species with the lowest representation of transposable elements. The highest level of intraspecific variability regarding their expression was found in *L. hargravesii* and *L. danicus*, where again one strain, 4B6, profoundly differed, and the lowest in *L. aporus.* The same was true for the related variants. This pattern following the general pattern of all transcripts indicates once more the linkage of the variants with the expression values. The differentiation in TEs that followed the same discrimination of species based on whole transcriptomes supports the notion that TEs play an important role in diatoms and could be involved in the adaptive evolution of species (Maumus et al., 2009). Different composition of transposons could tune the different properties for each species.

4.4.6. Conclusion

The comparative transcriptomics analysis on species of the genus *Leptocylindrus* revealed a high functional diversity among species, which in most cases seems to be related to the seasonal distribution patterns rather than to phylogenetic relationships as currently understood. Accordingly, the two species widely distributed across the year and abundant, namely, *L. aporus* and *L. danicus*, show high intraspecific diversity, both in terms of polymorphism variants and expression level, whereas *L. hargravesii* and *L. convexus* sit on the other end. The variant calling matches the DE analysis, as well as the related annotation which means that the variation at coding regions coming from SNPs and indels is likely to form the main genetic background to phenotypic variation (Morin et al., 2004). Beyond the hypothesis of the correlation of the observed variability with the species seasonal distribution, the larger populations of *L. aporus* and *L. danicus* in GoN might have also offered to these species a higher possibility to accumulate micro-variations, which led to the higher functional diversity. The plasticity of the species could be

a result of divergent proteins belonging to related gene families. However, the differential regulation of separate components of the same metabolic pathway, which results from post-transcriptional and post-translational modification for selected genes or proteins, respectively, could also be an important mechanism.

Another interesting result is the differences noted between the two cryptic and phylogenetically closest species, *L. danicus* and *L. hargravesii,* which otherwise are quite diverse regarding their seasonal distribution. This in particular but also the rest of the results are an important step towards the direction of using expression profiles to understand mechanisms underlying the distribution of species. The expression patterns of the species can provide indications about (a) the temporal and spatial ranges allowed by their intraspecific diversity and (b) their environmental requirements through specific functional traits that differ across species as to match the environmental conditions of each season. TEs and the HSF regulatory system might play critical role in the specific reactions and capabilities of each species in the many different stressful environments that they experience during the year.

Finally, the results of this chapter show that intraspecific functional variability can be as high as interspecific functional changes or variability resulting from strong perturbations. Therefore intraspecific variability is a very important factor influencing the response of species to different environments. The current study is one of the first studies that focused on the variability of gene expression in different strains in such a detail in phytoplankton and we consider that the results urge for a reconsideration of the way on which most studies investigate functional profile of species focusing on a single strain.

Chapter 5. Diversity and distribution of Leptocylindraceae: a DNA-metabarcoding approach

5.1. Introduction

A very important aspect in order to better understand the ecology and evolution of a species is the estimation of its actual diversity and species distribution in the natural environment. However this becomes a quite challenging task when it concerns unicellular eukaryotic organisms (protists). The microscope has been the main instrument used to discover the richness of the microbial world since it was invented 350 years ago. The few morphological features available for species differentiation, morphological stasis but also phenotypic plasticity have an impact on the reliability of taxonomic assignment by microscopy. It is only in the past 30 years that the identification of unicellular eukaryotes, as well as the way we understand microbial communities and their role in structuring ocean biogeochemical dynamics, has truly changed and that is thanks to the major development of DNA sequencing (Heidelberg et al., 2010). Molecular approaches have offered a more objective view of the microbial world. Indeed, most of the molecular-based revisions on the diversity of microorganisms have led to a better definition of the relevant taxonomic units and have revealed cryptic or pseudo-cryptic species in several taxa such as the diatoms *Chaetoceros, Skeletonema, Pseudo-nitzschia* and *Leptocylindrus* (Sarno et al., 2005; Mann and Vanormelingen, 2013; Degerlund et al., 2012, Lundholm et al., 2012; Nanjappa et al., 2013). In addition to the molecular identification of individual species in cultured strains, the sequencing of molecular markers (meta-barcoding) and of whole genomes (metagenomics) and transcriptomes (metatranscriptomics) recovered directly from environmental samples have offered the potential to address the diversity, structure and function of microbial communities with unprecedented resolution, thus starting a new era in Marine Ecology.

Long before the development of sequencing technologies, such as High Throughput Sequencing (HTS), studies based on the analysis of environmental DNA showed that molecular markers are a valuable tool for describing the spatial and temporal variations of eukaryotes (Diez et al., 2001; Moon-van der Staay et al., 2001; Edgcomb et al., 2011). The coupling of HTS and metabarcoding approach (further explained in 5.1.1) has proven to be an even more powerful tool providing further insights in microbial biodiversity with remarkable detail (Zinger et al., 2012; Georges et al.,

2014; Zimmermann et al., 2015). Large scale projects such as the European coastal survey BioMarKs (Logares et al., 2014; Massana et al., 2015), the global Tara Ocean project (de Vargas et al., 2015; Malviya et al., 2016) and the ongoing Ocean Sampling Day initiative (Kopf et al., 2015) are based on HTS metabarcoding. Similarly, the same method could be useful for the investigation of the seasonality and interannual variability of marine microbial communities in order to understand their relationship with environmental variations such as temperature, light availability and rainfall changes (Giovannoni and Vergin, 2012; Genitsaris et al., 2015; Piredda et al., 2016).

Protists live in almost any environment that contains water and many of them, such as algae, are photosynthetic, thus vital primary producers, especially in the ocean as part of the plankton. Due to a presumed easy dispersal by wind and water these microscopic organisms were believed to have a cosmopolitan distribution (Fenchel and Finlay, 2004). Environmental DNA extraction and the produced genetic data have indicated geographical structure in marine and freshwater species similar to those found in animals and vascular plants but of wider ranges and rarer local endemism (Foissner, 2008). The moderate endemicity model of Foissner assumes that one third of the microscopic organisms are morphological and/or genetic endemics, which supports the notion that we only know very little, about 20%, of the actual diversity in many protist groups. It was the ignorance of the low extinction rates over geological time and of the possibilities protists have to speciate due to their short generation time that stopped us from seeing this truth earlier (Foissner, 2008).

In diatoms as in other organisms, the different spatial, but also seasonal, range of each species can be related to their different biochemical (Nanjappa et al., 2014b) and physiological characteristics (Degerlund et al., 2012; Huseby et al., 2012). As already mentioned, the majority of molecular-based taxonomic revisions in diatoms have indicated the consideration of the presence of cryptic or pseudo-cryptic species while few cases of phenotypic plasticity have been identified (e.g. Shirokawa et al., 2012). The presence of cryptic species or even populations within species that alternate along seasons or show particular spatial distribution might be a result of adaptation to specific ecological and seasonal niches (Hendry and Day, 2005; Rynearson et al., 2006). Therefore,

spatial and temporal distribution can be considered a reflection of adaptation of species to the different environmental conditions, including temperature, a relationship that has been already explored for Leptocylindraceae in the previous chapters.

5.1.1. DNA metabarcoding

In order to overcome the difficulties and doubts of morphological identification mentioned above, recent biodiversity and distribution assessments tend to use DNA metabarcoding, i.e. molecular identification based on marker sequences, as their main method or a complementary one (Taberlet et al., 2012; Diaz-Real et al., 2015; Reva et al., 2015; Geisen et al., 2015). It must be clarified that metabarcoding is different from metagenomics since the goal of the first is to identify taxa while the latter focuses on the sequence-based analysis of collective genomes contained in an environmental sample (Riesenfeld et al., 2004).

Originally, DNA barcoding used standardized DNA barcodes, such as a region of the mitochondrial cytochrome c oxidase I gene (COI) for animals and of the large subunit of ribulose 1,5-bisphosphate carboxylase gene (*rbc*L) for plants, in order to identify species from more or less intact DNA isolated from single specimens using Sanger sequencing (Taberlet et al., 2012). DNA metabarcoding studies were initially performed through Sanger-sequencing of clone libraries (Stoeck and Epstein, 2003). Before sequencing, DNA fragments are combined with vector DNA (e.g. plasmids) to generate recombinant DNA molecules which are then introduced into a host organism, usually *E. coli* bacteria. The recombinant DNA molecules are replicated along with the host DNA resulting in a large number of clones of genetically modified bacteria. Recombinant DNA of each colony is isolated and sequenced. The read length can be quite high, up to 1,000 bp and the per-base accuracy as high as 99.999% (Shendure and Hanlee, 2008; Dewey et al., 2012). In GoN, metabarcoding studies using clone libraries have led to the assessment of molecular diversity and seasonality in the diatom genus *Pseudo-nitzschia* (McDonald et al., 2007b; Ruggiero et al., 2015) and molecular diversity in photosynthetic ultraplankton (McDonald et al., 2007a). However, Sanger is limited to sequencing a single gene from a single specimen in each run while high-throughput sequencers can separately sequence DNA molecules from a mixture of genes,

specimens and species. Massively parallel DNA sequencing platforms are now widely available, reducing the cost of sequencing by over two orders of magnitude and allowing individual investigators to use the sequencing capacity of a major genome center. High Throughput Sequencing (HTS) is a cyclic-array sequencing where PCR amplicons derived from a single library molecule end up spatially clustered on a planar surface and followed by alternative cycles of sequencing by synthesis and imaging of the full array at each cycle. A contiguous sequencing read for each array feature is built up by these successive iterations of enzymatic interrogation and imaging (Shendure and Hanlee, 2008; Fig.6.1.b). The 454 (454 Life Sciences/Roche) instrument was the first HTS platform but it is now abandoned due to the development of more advanced platforms in terms of amount of data produced, time needed and base-call error, such as HiSeq and MiSeq (http://www.illumina.com), Ion Personal Genome Machine and Ion Proton (http://www.lifetechnologies.com/us/en/home/brands/ion-torrent.html) and SOLiD Genome Sequencer (http://www.lifetechnologies.com). Further advancements in the technology are leading to new methods such as single molecule sequencing which do not require DNA amplification step, with platforms such as HeliScope Genome Analyser (http://www.helicosbio.com) and PacBioRSII (http://www.pacificbiosciences.com).

In DNA-metabarcoding specific nucleotide markers are PCR-amplified from environmental DNA using universal primers and then sequenced in a high-throughput sequencer producing a wide dataset of DNA sequences (Yinqiu et al., 2013). In the end it is possible to quickly trace organisms directly in their environment without the need for isolation and cultivation, offering also the ability of discovering new taxa and overcoming the bias against rare taxa in the case of deep sequencing. Two widely used markers are V4 (390 bp) and V9 (130 bp) of the 18s rRNA gene, typical targets in microbial eukaryotes and the best universal ones since 18s rRNA is present across all eukaryotes; it is also present in most public reference databases and generalist primers are available (Amaral-Zettler et al., 2006; Guillou et al., 2012; Decelle et al., 2014).

HTS sequencing offers important advantages relative to Sanger such as the potential of hundreds of millions of sequencing reads obtained in parallel and the low cost for DNA sequence

production. But of course HTS and metagenomic approaches are not perfect, not yet at least since these technologies are constantly improved. There are problems still remaining to be resolved such as the PCR-introduced errors during amplification, the much shorter reads compared to conventional sequencing, errors due to degradation of template DNA or errors during sequencing (new platforms are at least tenfold less accurate than Sanger sequencing) and the dependence of the results on the sequence coverage. In addition, due to high copy numbers of the target gene, some organisms could be overrepresented and their abundance overestimated (Heidelberg et al., 2010; Taberlet et al., 2012; Shendure and Hanlee, 2008). For Illumina, one of the most accurate sequencing platforms, errors at ≤0.1% is achieved for ≥75-85% of bases (Schirmer et al., 2016; Ross et al., 2013; Glenn, 2011). Sequencing errors can artificially inflate diversity estimates and thus, common practices such as sequence clustering and singleton removal, are used in order to avoid these artificial overestimates, although they may hide actual rare diversity. In addition, preferential PCR amplification or high copy numbers of the target gene of specific taxa, as mentioned above, restrict the possibility to extract quantitative information from HTS data. It must also be noted that the diversity deciphered from DNA metabarcoding depends to a great extent on the marker used in the analysis since each marker might have its own evolutionary rate in each organism and therefore a specific marker might be more suitable for one genus over another. Finally, wide and controlled reference datasets are needed, including sequences from properly identified individual species, to be able to reliably assign sequences to named taxa, but the coverage of these database is currently far to be complete. For example, 18S V9 information is available for only 159 of the 900 diatom genera (fossils included) listed by Fourtanier and Kociolek (1999).

5.1.2. Leptocylindraceae family diversity and distribution

Leptocylindraceae family is a good case study to test the ability of HTS metabarcoding to detect its real distribution and diversity, especially since the quite uniform morphology (narrow, cylindrical cells) makes it hard to identify each single species with certainty under the microscope. The diatoms of the *Leptocylindrus* genus are centric diatoms, common in the marine plankton

worldwide. Molecular phylogenies have resolved Leptocylindraceae at a basal position in the radial centrics (group of diatoms with the most ancient fossil record).The morphological similarities in combination with the considerable nucleotide differences, among the *Leptocylindrus* species contrast with far smaller nucleotide differences among species within diatom genera that are more diverse morphologically. These contrasting evidences led Nanjappa et al. (2013) to conclude that the genus *Leptocylindrus* is an ancient one containing just a few genetically distinct remnant species of a once far more diverse lineage. Alternatively, the marker regions of the *Leptocylindrus* species may have evolved particularly fast. According to a recent DNA metabarcoding study based on HTS at six coastal sites in European coastal waters, individual *Leptocylindrus* species were found to be widespread but not all of them were recorded everywhere (Fig. 5.1.2.1, Nanjappa et al., 2014a).

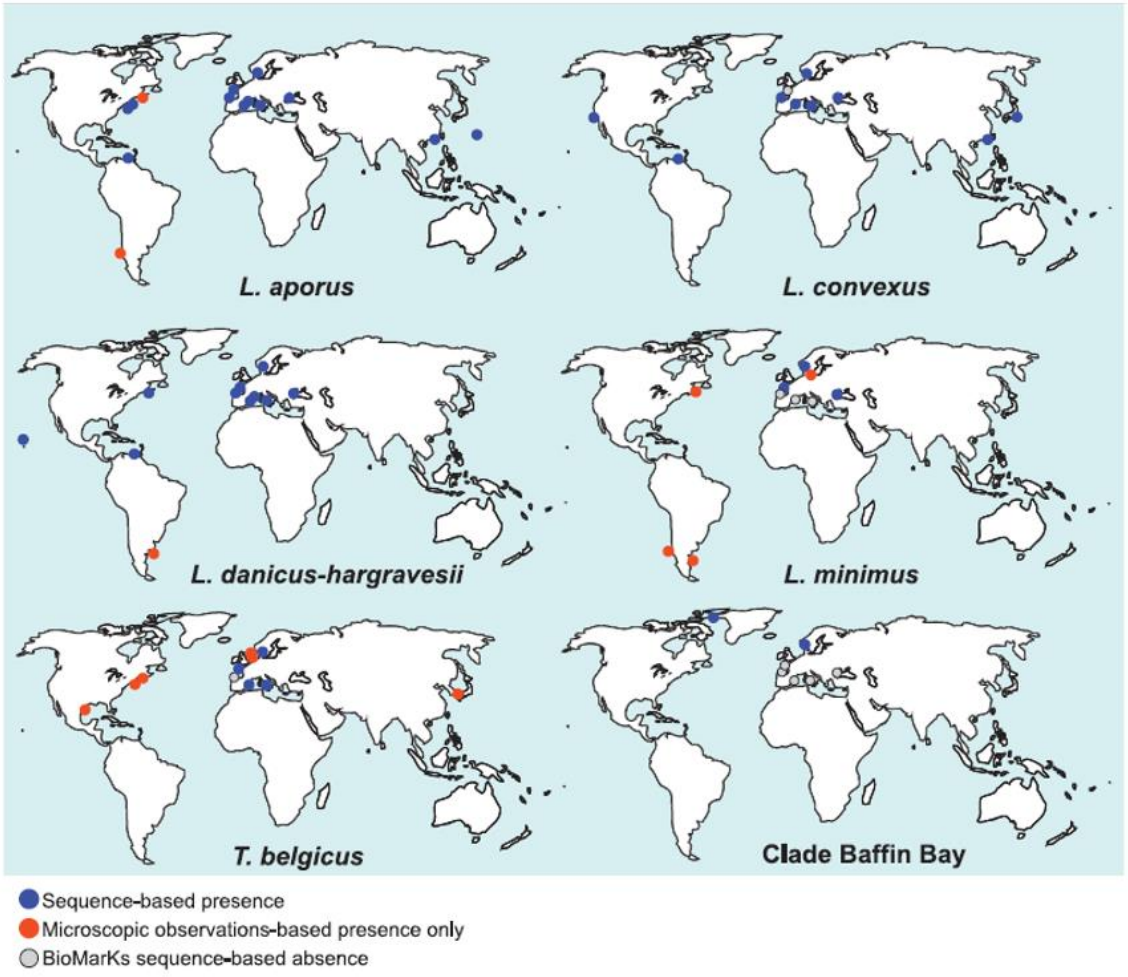


Figure 5.1.2.1. Distribution maps of Leptocylindraceae species inferred from HTS V4 and V9 sequences in the BioMarKs and GenBank datasets (blue dots), plus reliable microscopy images (red dots). Absence of finding in the BioMarKs dataset is represented by grey dots (Nanjappa et al., 2014a).

In that same study (Nanjappa et al., 2014a), both V4 (454 sequencing) and V9 (Illumina sequencing) regions of the nuclear-encoded SSU rDNA gene region of the BioMarKs project were analysed, as well as environmental sequences from the NCBI database. *Leptocylindrus aporus, L. danicus/ hargravesii, L. convexus* and *T. belgicus* sequences were found in the Mediterranean Sea, North Atlantic Ocean and Black Sea as well as at locations outside Europe while *L. minimus* sequences were found in the North Atlantic Ocean and the Black Sea but not in the Mediterranean Sea. In addition, V4 sequences belonging to a yet undescribed taxon were encountered only in Oslo Fjord and Baffin Bay (Nanjappa et al., 2014a).

An interesting fact in the case of Leptocylindraceae is that, despite simple morphology retained by all species, they are more distantly related compared to species in the genera *Skeletonema* and *Pseudo-nitzschia*, which in addition show comparatively higher morphological differentiation (Nanjappa et al., 2013). Furthermore, the known seasonal distribution based on isolation of strains from different seasons (Nanjappa et al., 2013) is quite diverse among species, which show different periods of high abundance. *Leptocylindrus aporus* was identified based on strains isolated from GoN from mid-July to mid-November, *L. danicus* was recorded from mid-November to mid-July, *L. hargravesii* was found only in December and January, and *L. convexus* was found in GoN from end of November to end of July. *Leptocylindrus minimus* was not found in GoN, instead it was *Tenuicylindrus belgicus* that had been identified under this name. *Tenuicylindrus belgicus* was found from the end of August to the beginning of November. All the information on the seasonal distribution are mainly based on Nanjappa's (2013) observations on cultivated strains from isolations that were performed almost weekly for more than a year. Despite the high extent of this attempt of seasonal reconstruction, there is always the possibility that a species escaped isolation during a specific season due to its rarity. As mentioned above, there has already been a study (Nanjappa et al., 2014a) on the assessment of the diversity and distribution of these species in European Seas using V4 and V9 as DNA metabarcoding markers. The data came from the project "Biodiversity of Marine Eukaryotes" (BioMarKs), a European Union ERA project involving experts from eight EU research institutes and aimed at studying eukaryotic microbial taxonomy

and evolution, genomics and molecular biology, marine biology and ecology, bioinformatics, as well as marine economy and policy. Nanjappa (2012) also explored the Tara Oceans dataset within the frames of his Ph.D. thesis but addressing only *Leptocylindrus* species as a preliminary overview. The "Tara Oceans expedition" was a global ocean expedition through 2009 - 2013 to study the impact of climate change on the microscopic life forms in the ocean including Arctic. Species diversity was recorded using the V9 variable region of the nuclear SSU rDNA. The main points that were concluded by Nanjappa and can be compared with the results of the present study are:

1. The species diversity in Leptocylindraceae is low, although there is intraspecific genetic diversity which would deserve further exploration.
2. Some species, such as *L. danicus* and *L. hargravesii*, are very similar when 18s rDNA is considered, and can be resolved in V4 but not in V9.
3. *L. aporus* sequences dominated the V4 BioMarKs and *L. danicus* the V9 BioMarKs dataset.
4. V4 produced more reliable distance trees compared to the lower resolution offered by the shorter V9 region.
5. V9 recovered many more sequences of Leptocylindraceae mainly due the higher sequencing depth used (Illumina sequencing versus 454 pyrosequencing), but in some cases due to a higher detection power of unknown origin.
6. In the BioMarKs dataset, among all sampling dates *L. aporus* dominated in October 2009 in the GoN while the rest of the species were also present on that date. In May 2010 all species were also present, though *L. danicus/L. hargravesii* and *L. convexus* were in higher numbers.

In the present chapter, the seasonal and spatial distribution of the Leptocylindraceae family was explored with the aim to discover any inter- or intraspecific diversity linked to specific temporal and/or spatial pattern. This investigation could help us better understand the strategies used by each species in order to cope with different environmental conditions and ultimately the ecology and evolution of the family. To this end, HTS DNA metabarcoding datasets from the Gulf of Naples

(GoN) (both V4 and V9 markers) and the BioMarKs and Tara stations (V9 marker) were analysed. What is particularly interesting in GoN is the option to compare the HTS metabarcoding data with the phytoplankton data collected in the ongoing time-series at Station MareChiara (40°48.5' N, 14°15' E). This dataset spans over the period 1984-2015 and contains data of phytoplankton species composition and abundance based on microscopy (Ribera d'Alcalà et al., 2004; Sarno and Zingone, 2008). The seasonal and worldwide distribution of Leptocylindraceae was assessed in a more extended way compared to previous efforts (Nanjappa, 2012; Nanjappa et al., 2014a) since HTS data from the GoN from several years and from all Tara stations (126 in contrast to only 35 Tara stations in Nanjappa 2012) were now available.

5.2. Materials and Methods

5.2.1. LTER MareChiara Dataset

As part of the analysis of environmental molecular data within Italian and European projects (EU-BioMarKs, FIRB Biodiversitalia) on protist biodiversity assessment in the GoN, HTS metabarcoding data were obtained on 48 different dates throughout three years (2011-2013) at the LTER-MC station (Roberta Piredda, personal communication). Every week, three liters of surface seawater were collected and three biological replicates were filtered on a cellulose ester filter (47 mm diameter, 1.2 µm pore size, EMD Millipore, USA). Filters were frozen immediately in liquid nitrogen and stored at -80 °C. DNA was extracted from each half of two filters collected at each sampling date using the DNeasy Plant Kit (QIAGEN GmbH, Hilden, Germany) following manufacturer's instructions by Dr. Roberta Piredda and Dr. Maria Paola Tomasino in Stazione Zoologica Napoli (SZN). DNA concentration and quality were measured with a NanoDrop Spectrophotometer (Thermo Fisher Scientific Inc, UK). V4 and V9, hypervariable regions of the small subunit (SSU) of rDNA, were used as targets for the high throughput sequencing. The following table shows the 48 dates of 2011-2013 from which environmental DNA has been sequenced (16 dates in 2011, 15 in 2012 and 17 in 2013):

Table 5.2.1.1 HTS metabarcoding sampling dates.

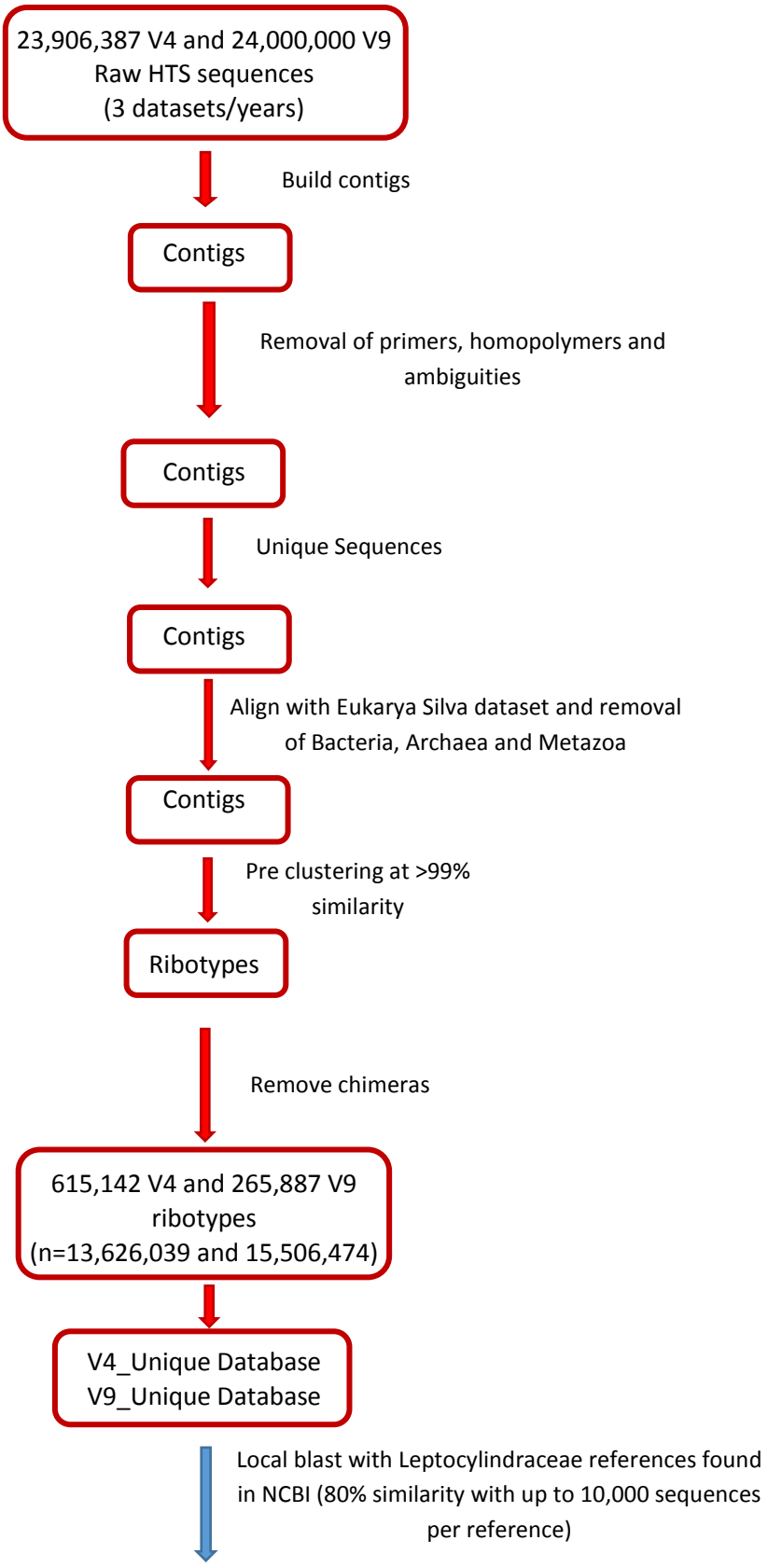
	2011	2012	2013
January	1st	17th	28th
February	15th	14th	19th
March	3rd	7th	28th
April	27th	3rd	16th and 30th
May	11th	4th	21st
June	7th and 21st	5th and 19th	4th and 18th
July	19th and 26th	10th and 31st	4th and 16th
August	16th and 30th	-	6th and 20th
September	6th and 27th	7th and 18th	10th
October	25th	2nd and 23rd	2nd and 28th
November	15th	13th	-
December	20th	23rd	4th and 30th

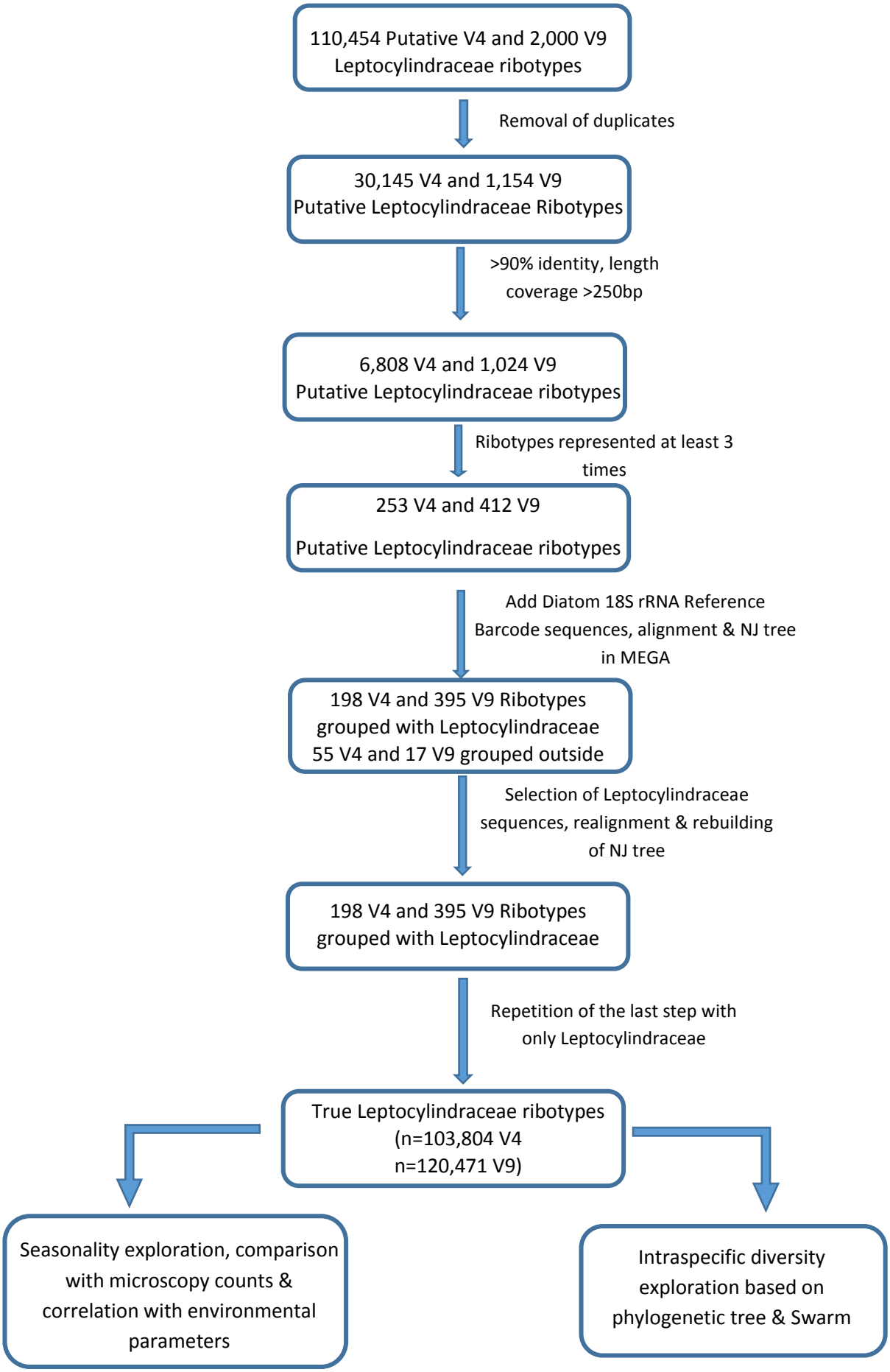
The sequencing was performed at the Molecular Biodiversity Lad (MoBiLab) of the ESFRI LifeWatch-Italy (CNR, Bari, Italy) on the Illumina MiSeq platform using BioMarKs primers (Stoeck et al., 2006) with slight modifications aimed at maximizing specificity for protists (Piredda et al., 2016). Two separate amplifications were performed on the DNA from the two half filters for each date and the PCR products were pooled in one sample per date. In the first amplification, V4 and V9 regions of 18S rRNA gene were amplified using the selected V4 and V9 universal primers having a 5' overhang sequence, corresponding to Nextera transposase primer (Piredda et al., 2016). Amplifications were performed in a reaction mixture containing 2.5 ng or 5 ng of extracted DNA (for V9 and V4 region, respectively), 1X Buffer HF, 0.2 mM dNTPs, 0.5 µM of each primer, and 1U of Phusion High-Fidelity DNA polymerase (New England Biolabs Inc, Ipswich, MA, USA) in a final volume of 25 µl. The cycling parameters for PCR were standardized as follows: initial denaturation 98 °C for 30 s, followed by 10 cycles of denaturation at 98 °C for 10 s, annealing at 44 or 56 °C (V4 and V9 region, respectively) for 30 s, extension at 72 °C for 15 s, and subsequently 15 cycles of denaturation at 98 °C for 10 s, annealing at 62 °C for 30 s, extension at 72 °C for 15 s, with a final extension step of 7 min at 72 °C. All PCRs were performed in presence of a negative control (RNase/DNase-free water). The PCR products (~270 bp for V9 and ~470 bp for V4) were visualized on 1.2% agarose gel and purified using the AMPure XP Beads (Agencourt Bioscience Corporation, Beverly, MA, USA), at a concentration of 1.2X vol/vol, according to manufacturer's instructions. The purified V4 and V9 amplicons were used as templates in the second PCR step, which was performed with the Nextera index primers and Illumina P5 and P7 primers as required by the

Nextera dual index approach. The 50 µl reaction mixture was made up of the following reagents: template DNA (40 ng), 1X Buffer HF, dNTPs (0.1 mM), Nextera index primers (index 1 and 2) and 1U Phusion DNA Polymerase. The cycling parameters were those suggested by the Illumina Nextera protocol. The final amplicon had a size of ca 550 bp (including ca 400bp of V4 region and 150 bp of Illumina Nextera adapters) and was purified using AMPure XP Beads, at a concentration of 0.6X vol/vol. The quality of the product was analysed on a 2100 Bioanalyzer (Agilent Technologies, Santa Clara, CA, USA) and quantified by fluorimetry using the Quant-iTTM PicoGreen-dsDNA Assay Kit (Thermo Fisher Scientific, Waltham, MA, USA) on NanoDrop 3300 (Thermo Fisher Scientific). Equimolar quantities of V4 and V9 amplicons were pooled and subjected to 2x250 bp sequencing on MiSeq platform to obtain a total of about 375,000 in V4 and 497,000 in V9 pair-end reads/sample.

The raw sequences were assessed for sequence quality using the FastQC tool on the Galaxy platform (http://usegalaxy.org/) before assembling. Illumina paired-end reads (ca. 250bp) were processed using Mothur v.1.33.0 (Schloss et al., 2009), using the standard operating procedure (http://www.mothur.org/wiki/MiSeq_SOP; Kozich et al., 2013). Initial HTS data included 48 surface water samples and two sediment samples. Sediment counts were excluded, as non-relevant to the thesis objectives. Pairs of forward and reverse sequences in the raw HTS data were aligned and merged to generate contigs. The entire V9 fragment was covered by two reads, whereas V4 overlap was on average 81 bp (St.dev. 11.3). Differences in base calls in the overlapping region were solved using the ΔQ parameter as described in Kozich et al. (2013). Primer sequences were removed and ambiguous bases were not allowed; homopolymers longer than 8bp were excluded. Dereplication and alignment to a reference alignment (silva.seed v119) followed for the remaining sequences. The ones that did not align to the target region were removed as well. Sequences were further denoised by the pre-clustering algorithm (Huse et al., 2010), allowing one nucleotide difference for every 100 bp of sequence, and the resulting sequences were screened for chimeras using UCHIME in *de novo* mode (Edgar et al., 2011). A naïve Bayesian classifier (Wang et al., 2007) trained using the PR2 database (Guillou et al., 2012) as training set with an 80% bootstrap

confidence threshold was performed in order to detect and remove Bacteria, Archaea and Metazoa. The whole process of data analysis, including pre-processing steps, is depicted in the following flowchart.





The initial steps (marked with red colour in the diagram above) of the raw data analysis were

performed by Roberta Piredda (including read statistics):

- HTS library preparation (contigs). Adapter trimming and quality control/ filtration (removal of homopolymers, chimeras and non-Silva aligned sequences).
- Pre-clustering at >99% similarity.

The specific procedure I used in my PhD thesis for the species of interest from that point on was the following (marked with blue colour in the diagram):

1. Leptocylindraceae sequences were extracted using local blast with reference sequences of all six species found in the National Center of Biotechnology Information (NCBI) (threshold of 80% identity and 10,000 sequences per query).
2. Duplicates were removed.
3. Sequences with at least 90% identity to the reference ones and a length coverage equal to or more than 250 base pairs for V4 and 120 base pairs for V9 were selected (V4 length is approximately 380 bp and V9 is approximately 120 base pairs).
4. In order to better refine the selectivity of our dataset, the ribotypes (unique sequences) that were represented less than three times in the dataset were eliminated. This specific number of times was chosen as a minimum threshold taking into consideration that a reliable ribotype in a study regarding the seasonality should appear at least once per year over the three study years.
5. The selected sequences together with the references of Leptocylindraceae but also of all known diatoms were aligned in MAFFT v.7, a multiple sequence alignment program for amino acid or nucleotide sequences (Katoh and Standley, 2013). The alignment was checked by eye for errors and edited if needed with the SeaView multiplatform graphical user interface for sequence alignment, v.4.5.2 (Gouy et al., 2010). No sequence was removed because of a diagnosed technical error.
6. Based on the alignment a phylogenetic tree was built in Molecular Evolutionary Genetics Analysis (MEGA) v.6 (Tamura et al., 2013), an integrated tool inferring phylogenetic trees. The statistical method used was neighbor-joining with bootstrap method set as test of

phylogeny (500 replications) and Kimura 2-parameter model as substitution model. All ambiguous positions were removed for each sequence pair.

7. HTS sequences that grouped with Leptocylindraceae reference sequences were selected, re-aligned with all diatom references and a NJ tree using the same settings was built again in order to double check their positioning on the tree. Sequences that grouped with references other than Leptocylindraceae were blasted in NCBI database and removed when confirmed to match other taxa.
8. HTS sequences that grouped with Leptocylindraceae reference sequences were selected, realigned only with Leptocylindraceae references and a NJ tree using the same settings was built again in order to check the intraspecific variability. In addition, a maximum likelihood analysis (with 500 bootstrap replications, a General Time Reversible model with gamma distribution and invariant sites [GTR+G+I]) was performed to identify phylogenetic relationships and monophyletic groups (clades). Sequences resolved outside a terminal clade containing reference sequences were blasted in NCBI database and either considered as false positive or treated as belonging to Leptocylindraceae. All presented trees were built in MEGA while *Chaetoceros dayaensis* was used as outgroup.
9. The final HTS sequences corresponding to each species were selected and their specific counts through the 48 dates were retrieved from the dataset using mothur, a widely used bioinformatics tool for analyzing gene sequences using command lines (Schloss et al., 2009). Thus, the distribution of the family based on environmental DNA was calculated, based on abundances and after being standardized to the total number of diatoms.
10. The seasonality of each species was explored and compared with the phytoplankton data collected at the Station MareChiara (courtesy of Diana Sarno). Regarding the Leptocylindraceae, there are no counts available for *L. hargravesii* since it is very similar to *L. danicus* and it is almost impossible to distinguish between these two species in light microscopy. The identification of *L. aporus* as distinct from *L. danicus/L. hargravesii*, which is mainly based on chloroplast shape and number, is also hard in LM on fixed material. In

addition, environmental parameters including temperature, salinity, CTD measurements (conductivity, temperature and depth of the ocean), nutrients and chlorophyll acquired from the LTER-MC Station for all three years (courtesy of MEDA service, SZN) and explored for any correlation/association with Leptocylindraceae seasonality (see section 5.2.3 for details).

11. The intraspecific diversity was explored in more details with *Swarm*, a robust and fast clustering method for amplicon-based studies (Mahè et al., 2014). The method was used in order to cluster the V4 and V9 sequences that were concluded to belong to Leptocylindraceae into molecular operational taxonomic units (OTUs). The default number of differences allowed between two amplicons (ribotypes) was 1 (*d*=1), so two amplicons were grouped together if they had 1 or no differences. The number of differences is calculated as the number of mismatches including substitutions, insertions or deletions. Each OTU produced was explored and compared with the previous phylogenetic results.

In cases of species detected in unexpected periods of the year, isolations of species from samples collected at the LTER-MC station and molecular identification were carried out in order to confirm the information obtained from molecular data (see Chapter 2 for isolation, cultivation and molecular identification methods).

5.2.2. Tara Dataset

The V9 Eukaryotic Diversity dataset (http://Taraoceans.sb-roscoff.fr/EukDiv/) provided by Shruti Malviya (personal communication) was explored for Leptocylindraceae. The dataset was generated from 1,086 plankton samples collected at 154 stations (Tara Oceans – 126; Tara Arctic – 20; BioMarKs - 8) from four major organismal size fractions – pico-nanoplankton (0.8 to 5 µm), nanoplankton (5 to 20 µm), microplankton (20 to 180 µm) and mesoplankton (180 to 2000 µm) – sampled at two water-column depths: subsurface mixed-layer waters and deep chlorophyll maximum (DCM) at the top of the thermocline. This dataset has ~6.3 million barcodes (cleaned) corresponding to ~1.8 billion reads. The steps followed in the analysis were the same as described above, starting from local blast with Leptocylindraceae references. In this case we included

ribotypes that were present in at least 2 samples and in at least 3 copies. The BioMarKs sequences were included for the molecular diversity analysis but for the spatial distribution analysis they were removed. In addition in the distribution analysis the size fraction (5-20 µm) with the highest Leptocylindraceae abundance was selected and the different depths were treated separately in order to keep a uniform dataset. The standardized abundances of the assigned sequences were then plotted on the world map using the R package rworldmap and function mapPies (R Core Team, 2012) as well as Plotly bubble map tool (https://plot.ly).

5.2.3. Relationships of Leptocylindraceae with environmental variables

In order to explore the possible relationships of Leptocylindraceae distribution with environmental variables, the dataset of chemical and physical variables collected at the LTER-MC (courtesy of MECA service) and for the Tara Oceans stations (courtesy of the Tara Consortium) were used. The environmental parameters available for the latter dataset did not include the Arctic stations, two stations in Mediterranean Sea and one North Atlantic station. The following analyses were performed:

1. Due to the very low number of sequences in some stations and dates compared to others, rarefaction was used in order to transform the datasets. Rarefaction approaches are used in ecology to evaluate sampling effort and community richness while it is also included as a normalization step in widely used microbial community analysis pipelines such as QIIME (Brewer and Williamson, 1994; Caporaso et al., 2010; Gotelli and Colwell, 2001). Function *rrarefy* of the vegan R package was used to normalize data; it gives the expected species richness of the community in random subsamples of equal size from each site. In this case, the size of the subsample was chosen to be equal to the median of each dataset.
2. Communities were clustered based on their compositional similarity. Hierarchical cluster analysis (HCA) and canonical correlation analysis (CCA) were used for this purpose.

- HCA is a clustering method which builds a hierarchy of clusters based on dissimilarity. The function *vegdist* and *hclust* of the R package vegan were used. Vegdist computes dissimilarity indices using quantitative data. As dissimilarity index, we used the Bray-

Curtis index which quantifies the compositional dissimilarity between two different sites (or dates) based on counts at each site. Bray-Curtis index delivers robust and reliable dissimilarity results for a wide range of applications and therefore is one of the most commonly applied measurements to express relationships in ecology and environmental sciences. It is also not affected by the number of null values between samples like Euclidean distance so it is preferred since our data have a high number of zeros. As defined by Bray and Curtis (1957), the index of dissimilarity is:

$BC_{ij} = 1 – (2C_{ij} / (S_i+S_j))$

where C_{ij} is the sum of the lesser values for only those species in common between both sites, S_i and S_j are the total number of specimens counted at both sites. The Bray-Curtis dissimilarity is bound between 0 and 1, where 0 means the two sites (or dates) have the same composition (they share all species) and 1 means exactly the opposite (they do not share any species). *hclust* function was applied on the distance matrix built by *vegdist* and plotted a dendrogram that displays the hierarchical relationship among the sites (dates). This approach is 'bottom up'. The cluster distance between two clusters is the maximum distance between their individual components (species counts). At every step of the process, the two nearest clusters are merged into a new one and this is repeated until the whole dataset forms one single cluster.

- CCA is a multivariate constrained ordination technique. Unconstrained ordination (as NMDS) tries to display all the variation in data while constrained ordination tries to display only the variation that can be explained with constraining variables. CCA is a good choice if there is a clear and strong a priori hypothesis on constraints and no interest in the major structure in the dataset. Therefore NMDS is used for exploration of data and CCA (*cca* function in R) for the environmental interpretation (Ramette, 2007).

3. In the final step, the data matrix of similarities (species counts in each station/date) is compared to the variables (environmental parameters) using the *bioenv* function of the vegan R package. The environmental variables are explored to find the best subset of them that correlate to sample similarities of the biological community (species abundance). In particular, *bioenv* aims to the Euclidean distances of scaled environmental variables that have the maximum (rank) correlation with community dissimilarities.

The Tara Leptocylindraceae dataset was treated as a whole in a first stage but specific fractions were selected for further analysis in order to compare the abundances among different stations in a homogeneous dataset. In addition, since the sampling in the different stations was performed on different dates, the possible effect of the sampling time period on the species abundances in each station was explored.

5.3. Results

5.3.1. Molecular Diversity

The V4 and V9 LTER-MC datasets

For V4, the preprocessing procedure gave 13,626,039 sequences starting from 23,906,389 raw HTS sequences. After preclustering 615,142 unique sequences made up the V4 database. The local blast against Leptocylindraceae produced 110,454 sequences of which 30,145 were unique. 6,808 sequences were equal or more than 90% identical to Leptocylindraceae references with length coverage equal or more than 250 bp. In the end 253 ribotypes represented 3 or more times in the dataset were used for the NJ tree with references of all diatoms. 198 ribotypes were grouped with Leptocylindraceae references and 55 ribotypes grouped with other diatoms on the phylogenetic tree.

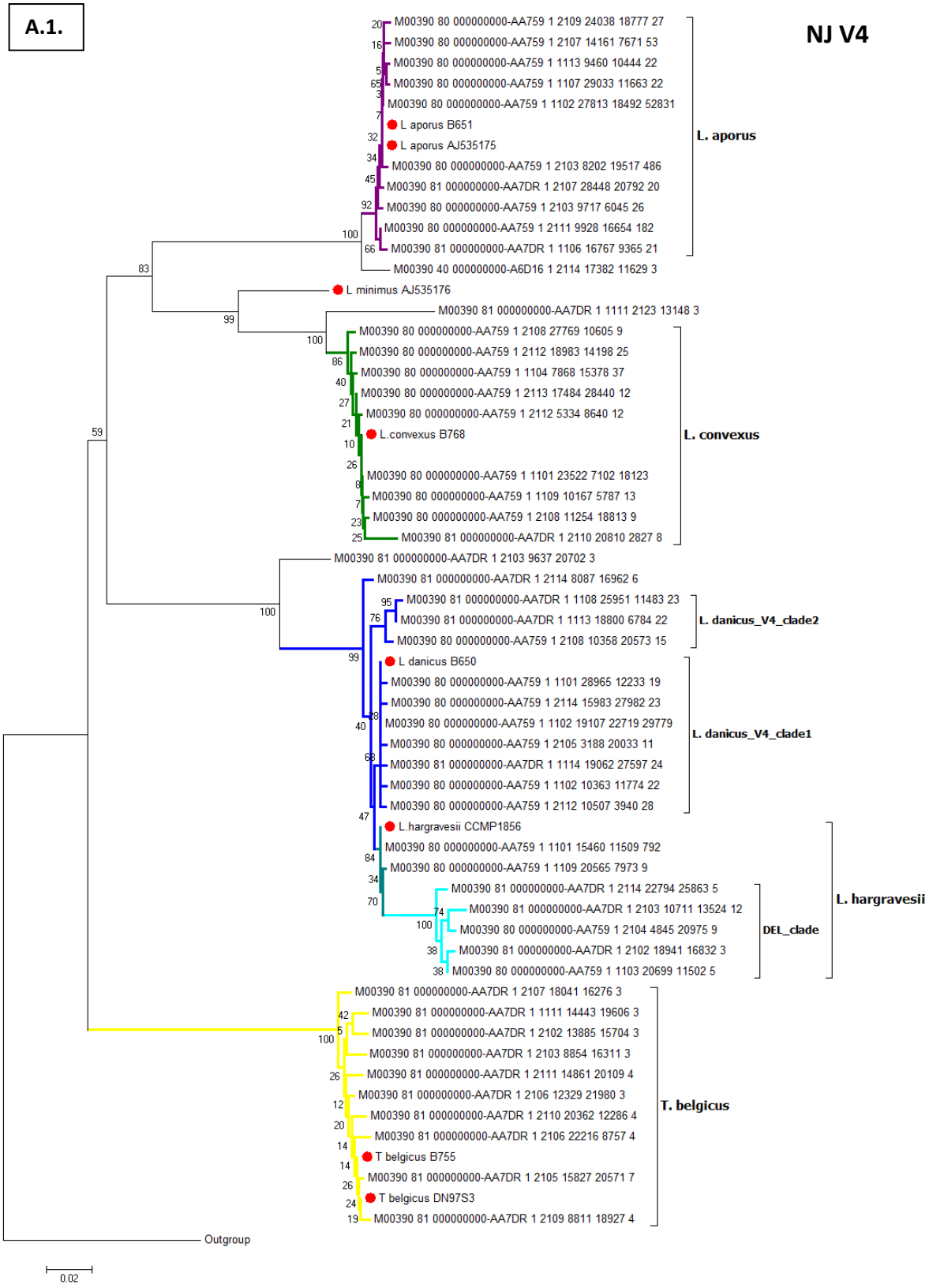
For V9, the preprocessing procedure gave 15,506,474 sequences starting from 24,000,000 raw HTS sequences. After preclustering 265,887 unique sequences made up the V9 database. The local blast against Leptocylindraceae produced 2,000 sequences of which 1,154 were unique. 1,024 sequences were equal or more than 90% identical to Leptocylindraceae references with a length coverage $\geq$ 120 bp. In the end 412 ribotypes represented 3 or more times in the dataset were

used for the NJ tree with references of all diatoms. 395 ribotypes were grouped with Leptocylindraceae references and 17 ribotypes grouped with other diatoms on the phylogenetic tree. Detailed description of the number of sequences assigned to each species follows (Table 5.3.1.1).

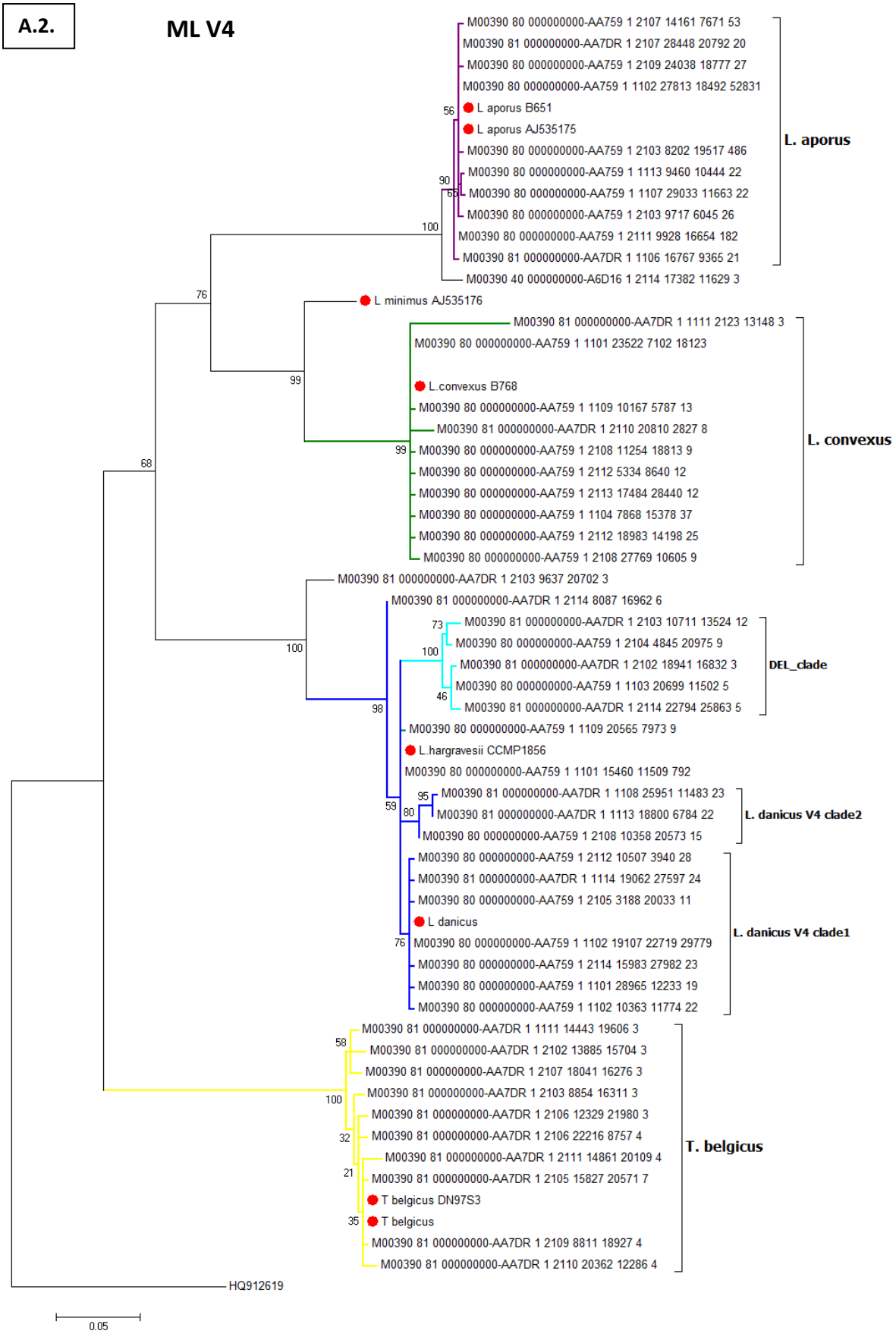
Table 5.3.1.1 Number of unique (ribotypes) and total sequences detected for each species based on V4 and V9 in LTER-MC dataset.

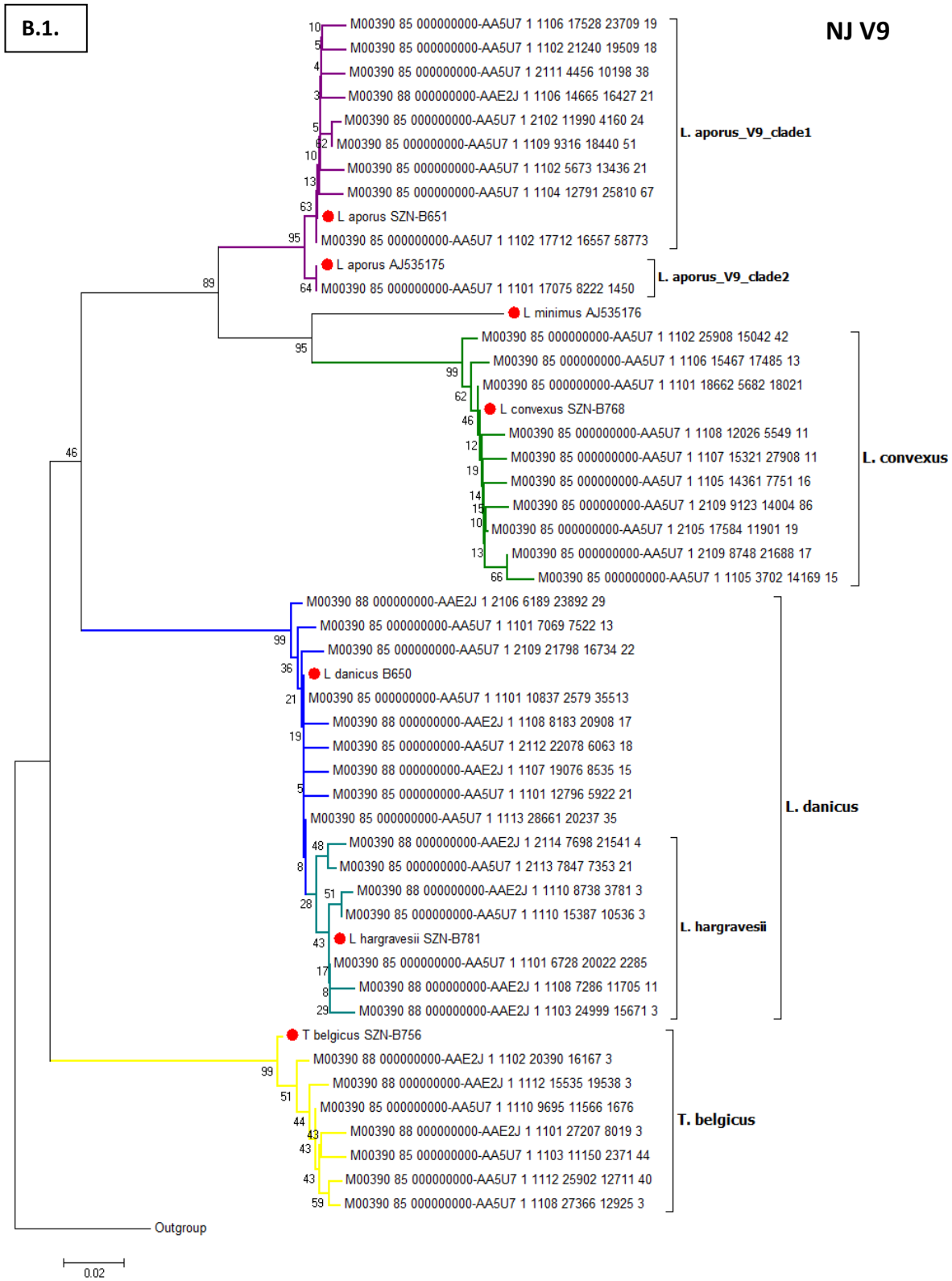
	V4		V9	
	N of ribotypes	**N of Seqs**	**N of ribotypes**	**N of Seqs**
L. aporus	98	54,406	168	61,531
L. danicus	53	30,204	117	36,185
L. hargravesii	8	841	7	2,312
T. belgicus	16	56	7	1,772
L. convexus	23	18,297	96	18,671
Total Number	198	103,804	395	120,471

No sequences matched *L. minimus* which has never been retrieved so far in the Gulf of Naples so this result was expected. All the phylogenetic trees showed the expected topology within the Leptocylindraceae family (Fig.5.3.1.1.A and B). NJ and ML trees were in accordance with each other regarding the assignment of sequences to species and clades.


A.1.
NJ V4
M00390 80 000000000-AA759 1 2109 24038 18777 27
M00390 80 000000000-AA759 1 2107 14161 7671 53
M00390 80 000000000-AA759 1 1113 9460 10444 22
M00390 80 000000000-AA759 1 1107 29033 11663 22
M00390 80 000000000-AA759 1 1102 27813 18492 52831
L aporus B651
L aporus AJ535175
M00390 80 000000000-AA759 1 2103 8202 19517 486
M00390 81 000000000-AA7DR 1 2107 28448 20792 20
M00390 80 000000000-AA759 1 2103 9717 6045 26
M00390 80 000000000-AA759 1 2111 9928 16654 182
M00390 81 000000000-AA7DR 1 1106 16767 9365 21
L. aporus
M00390 40 000000000-A6D16 1 2114 17382 11629 3
L minimus AJ535176
M00390 81 000000000-AA7DR 1 1111 2123 13148 3
M00390 80 000000000-AA759 1 2108 27769 10605 9
M00390 80 000000000-AA759 1 2112 18983 14198 25
M00390 80 000000000-AA759 1 1104 7868 15378 37
M00390 80 000000000-AA759 1 2113 17484 28440 12
M00390 80 000000000-AA759 1 2112 5334 8640 12
L.convexus B768
M00390 80 000000000-AA759 1 1101 23522 7102 18123
M00390 80 000000000-AA759 1 1109 10167 5787 13
M00390 80 000000000-AA759 1 2108 11254 18813 9
M00390 81 000000000-AA7DR 1 2110 20810 2827 8
L. convexus
M00390 81 000000000-AA7DR 1 2103 9637 20702 3
M00390 81 000000000-AA7DR 1 2114 8087 16962 6
M00390 81 000000000-AA7DR 1 1108 25951 11483 23
M00390 81 000000000-AA7DR 1 1113 18800 6784 22
M00390 80 000000000-AA759 1 2108 10358 20573 15
L. danicus_V4_clade2
L danicus B650
M00390 80 000000000-AA759 1 1101 28965 12233 19
M00390 80 000000000-AA759 1 2114 15983 27982 23
M00390 80 000000000-AA759 1 1102 19107 22719 29779
M00390 80 000000000-AA759 1 2105 3188 20033 11
M00390 81 000000000-AA7DR 1 1114 19062 27597 24
M00390 80 000000000-AA759 1 1102 10363 11774 22
M00390 80 000000000-AA759 1 2112 10507 3940 28
L. danicus_V4_clade1
L.hargravesii CCMP1856
M00390 80 000000000-AA759 1 1101 15460 11509 792
M00390 80 000000000-AA759 1 1109 20565 7973 9
M00390 81 000000000-AA7DR 1 2114 22794 25863 5
M00390 81 000000000-AA7DR 1 2103 10711 13524 12
M00390 80 000000000-AA759 1 2104 4845 20975 9
M00390 81 000000000-AA7DR 1 2102 18941 16832 3
M00390 80 000000000-AA759 1 1103 20699 11502 5
DEL_clade
L. hargravesii
M00390 81 000000000-AA7DR 1 2107 18041 16276 3
M00390 81 000000000-AA7DR 1 1111 14443 19606 3
M00390 81 000000000-AA7DR 1 2102 13885 15704 3
M00390 81 000000000-AA7DR 1 2103 8854 16311 3
M00390 81 000000000-AA7DR 1 2111 14861 20109 4
M00390 81 000000000-AA7DR 1 2106 12329 21980 3
M00390 81 000000000-AA7DR 1 2110 20362 12286 4
M00390 81 000000000-AA7DR 1 2106 22216 8757 4
T belgicus B755
M00390 81 000000000-AA7DR 1 2105 15827 20571 7
T belgicus DN97S3
M00390 81 000000000-AA7DR 1 2109 8811 18927 4
T. belgicus
Outgroup
0.02

A.2.


ML V4
M00390 80 000000000-AA759 1 2107 14161 7671 53
M00390 81 000000000-AA7DR 1 2107 28448 20792 20
M00390 80 000000000-AA759 1 2109 24038 18777 27
M00390 80 000000000-AA759 1 1102 27813 18492 52831
L aporus B651
L aporus AJ535175
M00390 80 000000000-AA759 1 2103 8202 19517 486
M00390 80 000000000-AA759 1 1113 9460 10444 22
M00390 80 000000000-AA759 1 1107 29033 11663 22
M00390 80 000000000-AA759 1 2103 9717 6045 26
M00390 80 000000000-AA759 1 2111 9928 16654 182
M00390 81 000000000-AA7DR 1 1106 16767 9365 21
M00390 40 000000000-A6D16 1 2114 17382 11629 3
L. aporus
L minimus AJ535176
M00390 81 000000000-AA7DR 1 1111 2123 13148 3
M00390 80 000000000-AA759 1 1101 23522 7102 18123
L.convexus B768
M00390 80 000000000-AA759 1 1109 10167 5787 13
M00390 81 000000000-AA7DR 1 2110 20810 2827 8
M00390 80 000000000-AA759 1 2108 11254 18813 9
M00390 80 000000000-AA759 1 2112 5334 8640 12
M00390 80 000000000-AA759 1 2113 17484 28440 12
M00390 80 000000000-AA759 1 1104 7868 15378 37
M00390 80 000000000-AA759 1 2112 18983 14198 25
M00390 80 000000000-AA759 1 2108 27769 10605 9
L. convexus
M00390 81 000000000-AA7DR 1 2103 9637 20702 3
M00390 81 000000000-AA7DR 1 2114 8087 16962 6
M00390 81 000000000-AA7DR 1 2103 10711 13524 12
M00390 80 000000000-AA759 1 2104 4845 20975 9
M00390 81 000000000-AA7DR 1 2102 18941 16832 3
M00390 80 000000000-AA759 1 1103 20699 11502 5
M00390 81 000000000-AA7DR 1 2114 22794 25863 5
DEL_clade
M00390 80 000000000-AA759 1 1109 20565 7973 9
L.hargravesii CCMP1856
M00390 80 000000000-AA759 1 1101 15460 11509 792
M00390 81 000000000-AA7DR 1 1108 25951 11483 23
M00390 81 000000000-AA7DR 1 1113 18800 6784 22
M00390 80 000000000-AA759 1 2108 10358 20573 15
L. danicus V4 clade2
M00390 80 000000000-AA759 1 2112 10507 3940 28
M00390 81 000000000-AA7DR 1 1114 19062 27597 24
M00390 80 000000000-AA759 1 2105 3188 20033 11
L danicus
M00390 80 000000000-AA759 1 1102 19107 22719 29779
M00390 80 000000000-AA759 1 2114 15983 27982 23
M00390 80 000000000-AA759 1 1101 28965 12233 19
M00390 80 000000000-AA759 1 1102 10363 11774 22
L. danicus V4 clade1
M00390 81 000000000-AA7DR 1 1111 14443 19606 3
M00390 81 000000000-AA7DR 1 2102 13885 15704 3
M00390 81 000000000-AA7DR 1 2107 18041 16276 3
M00390 81 000000000-AA7DR 1 2103 8854 16311 3
M00390 81 000000000-AA7DR 1 2106 12329 21980 3
M00390 81 000000000-AA7DR 1 2106 22216 8757 4
M00390 81 000000000-AA7DR 1 2111 14861 20109 4
M00390 81 000000000-AA7DR 1 2105 15827 20571 7
T belgicus DN97S3
T belgicus
M00390 81 000000000-AA7DR 1 2109 8811 18927 4
M00390 81 000000000-AA7DR 1 2110 20362 12286 4
T. belgicus
HQ912619
0.05


B.1.
NJ V9
M00390 85 000000000-AA5U7 1 1106 17528 23709 19
M00390 85 000000000-AA5U7 1 1102 21240 19509 18
M00390 85 000000000-AA5U7 1 2111 4456 10198 38
M00390 88 000000000-AAE2J 1 1106 14665 16427 21
M00390 85 000000000-AA5U7 1 2102 11990 4160 24
M00390 85 000000000-AA5U7 1 1109 9316 18440 51
M00390 85 000000000-AA5U7 1 1102 5673 13436 21
M00390 85 000000000-AA5U7 1 1104 12791 25810 67
L aporus SZN-B651
M00390 85 000000000-AA5U7 1 1102 17712 16557 58773
L. aporus_V9_clade1
L aporus AJ535175
M00390 85 000000000-AA5U7 1 1101 17075 8222 1450
L. aporus_V9_clade2
L minimus AJ535176
M00390 85 000000000-AA5U7 1 1102 25908 15042 42
M00390 85 000000000-AA5U7 1 1106 15467 17485 13
M00390 85 000000000-AA5U7 1 1101 18662 5682 18021
L convexus SZN-B768
M00390 85 000000000-AA5U7 1 1108 12026 5549 11
M00390 85 000000000-AA5U7 1 1107 15321 27908 11
M00390 85 000000000-AA5U7 1 1105 14361 7751 16
M00390 85 000000000-AA5U7 1 2109 9123 14004 86
M00390 85 000000000-AA5U7 1 2105 17584 11901 19
M00390 85 000000000-AA5U7 1 2109 8748 21688 17
M00390 85 000000000-AA5U7 1 1105 3702 14169 15
L. convexus
M00390 88 000000000-AAE2J 1 2106 6189 23892 29
M00390 85 000000000-AA5U7 1 1101 7069 7522 13
M00390 85 000000000-AA5U7 1 2109 21798 16734 22
L danicus B650
M00390 85 000000000-AA5U7 1 1101 10837 2579 35513
M00390 88 000000000-AAE2J 1 1108 8183 20908 17
M00390 85 000000000-AA5U7 1 2112 22078 6063 18
M00390 88 000000000-AAE2J 1 1107 19076 8535 15
M00390 85 000000000-AA5U7 1 1101 12796 5922 21
M00390 85 000000000-AA5U7 1 1113 28661 20237 35
M00390 88 000000000-AAE2J 1 2114 7698 21541 4
M00390 85 000000000-AA5U7 1 2113 7847 7353 21
M00390 88 000000000-AAE2J 1 1110 8738 3781 3
M00390 85 000000000-AA5U7 1 1110 15387 10536 3
L hargravesii SZN-B781
M00390 85 000000000-AA5U7 1 1101 6728 20022 2285
M00390 88 000000000-AAE2J 1 1108 7286 11705 11
M00390 88 000000000-AAE2J 1 1103 24999 15671 3
L. hargravesii
L. danicus
T belgicus SZN-B756
M00390 88 000000000-AAE2J 1 1102 20390 16167 3
M00390 88 000000000-AAE2J 1 1112 15535 19538 3
M00390 85 000000000-AA5U7 1 1110 9695 11566 1676
M00390 88 000000000-AAE2J 1 1101 27207 8019 3
M00390 85 000000000-AA5U7 1 1103 11150 2371 44
M00390 85 000000000-AA5U7 1 1112 25902 12711 40
M00390 85 000000000-AA5U7 1 1108 27366 12925 3
T. belgicus
Outgroup
0.02

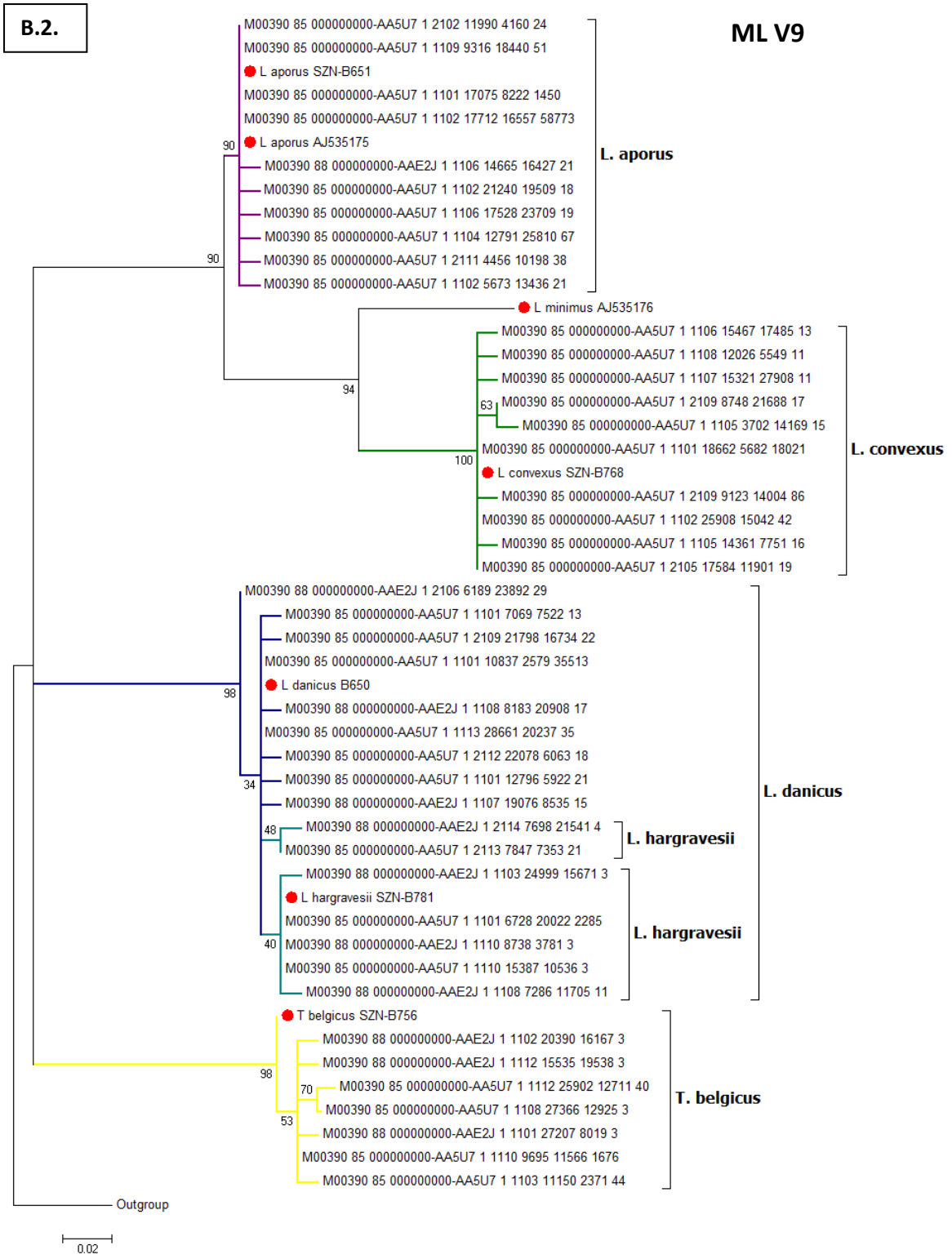


Figure 5.3.1.1 Neighbor-joining with Kimura 2-parameter model as substitution model (1) and maximum-likelihood with GTR substitution model (2) trees based on HTS V4 (A) and V9 (B) Leptocylindraceae sequences in the period 2011-2013 at LTER-MC station, with bootstrap method set as test of phylogeny (500 replications). For a clear representation of the tree here and only for this, the ten most abundant ribotypes of each species were selected. The last number in the ribotype labels represents the number of sequences.

In V4, three main clades were detected: the first one included the reference sequences of *L. aporus, L. minimus* and *L. convexus*; the second clade included *L. danicus* and *L. hargravesii* and

the third *Tenuicylindrus belgicus.* The most variable species was *L. danicus*, showing the presence of a distinct subclade (clade 2).

In the *L. aporus* V4 clade, the most numerous sequences were identical to the references and they were all grouped with the two reference sequences except one (M00390_40_000000000-A6D16_1_2114_17382_11629), which when blasted returned *L. aporus* with 98% identity (8 mismatches) and was represented by three individuals; all found on September 27th, 2011. After careful check of the alignment, the sequence was not considered to be product of an artificial error since mismatches were present in variable sites among Leptocylindraceae and other diatoms.

In the *L. convexus* V4 clade, the most numerous sequences were identical to the reference sequences. All ribotypes grouped with the references except one (M00390_81_000000000-AA7DR_1_1111_2123_13148) which when blasted returned *L. convexus* with 95% identity (21 mismatches) and was represented by three individuals; all found on 18 June 2013. After careful check of the alignment, the sequence could not be concluded to be product of an artificial error with certainty since mismatches were present both in variable and conserved sites among Leptocylindraceae and other diatoms.

In *L. danicus* V4, two distinct clades were retrieved, one with most sequences including the reference one and the other with 70 sequences (*L. danicus* V4 clade2) with 98% identity to the reference sequence, 3 - 7 mismatches. Finally, one sequence (M00390_81_000000000-AA7DR_1_2103_9637_20702) fell totally outside the *L. danicus* clade, matching the reference at 94% (23 mismatches) and represented by three sequences detected on one date (July 16th 2013). After careful check of the alignment, the sequence could not be concluded to be product of an artificial error with certainty since mismatches were present both in variable and conserved sites among Leptocylindraceae and other diatoms. In the *L. hargravesii* clade most sequences were identical or very similar to the reference but a separate sub-clade was also found with a 95% identity. The alignment revealed 19 mismatches with the reference sequence, including a deletion of 7 nucleotides. These ribotypes (*L. hargravesii* V4 DEL clade) were present on several dates in

2011 and more extensively in 2013.

Finally, the *T. belgicus* clade was relatively homogeneous despite the small difference between the two reference sequences.

The V9 tree presented the same three main clades as the V4 tree but with some differences within species. Here the most variable species was *L. aporus* with the presence of a distinct subclade. In particular, the *L. aporus* clade was well supported but the two references grouped separately. The two *L. aporus* references show one nucleotide mismatch, enough to differentiate the small population consisting of four ribotypes (only one shown in the representative tree of Fig. 5.3.1.1) which were more similar to the sequence AJ535175, from an Atlantic strain isolated by Hargraves in 1986. The fact that the nucleotide mismatch is within a site that is highly conserved in *L. aporus,* but variable compared to other Leptocylindraceae, supports the view that the variants are real. Yet again, the *L. aporus* V9 clade 2 is obvious in the NJ tree and the Swarm analysis (see below) but the signal is lost in the ML tree due to the small genetic distance between the ribotypes. *L. danicus* and *L. hargravesii* in V9 were not well separated, as they show only one nucleotide difference, and there was no other clade detected within these two species. *Tenuicylindrus belgicus* was again rather homogeneous, although most sequences slightly differed from the reference one.

For a better representation and interpretation especially of the intraspecific diversity, swarm analysis was also used. Swarm analysis is a fast and robust method based on a maximum number of differences *d* and focuses only on very close relationships. It has been shown to work much better for large datasets, such as BioMarKs and Tara, compared to other greedy *de novo* clustering methods. It recursively groups amplicons with *d* or less differences and produces stable clusters (or “swarms”). Therefore, swarm analysis can give a better idea of the interspecific relationships and the clades within each species are detected in a more straightforward way than in phylogenetic trees when the number of sequences is high. The graphical representation of the swarm-derived OTUs assists to this end. Despite being less informative for the smaller MC dataset, it was also used in this case in order to keep a uniform analysis for both datasets which could also

lead to a better comparison of results.

The results of the swarm clustering can be seen in the table below:

Table 5.3.1.2 Number of total OTUs produced and OTUs consisting of only one sequence.

Marker	Total OTUs	OTUs consisting of one sequence
V4	100	93
V9	12	6

The OTUs consisting of more than one representative sequence were 7 for V4 and 6 for V9. The following table presents some more detailed statistics on these OTUs:

Table 5.3.1.3 Statistics on each OTU for V4 and V9 at LTER-MC station. OTUs with only one ribotype are not presented. The seed is the representative sequence of each OTU which is also the most abundant one. The numbering of the OTUs depends on their abundance; so OTU#6 was more abundant than OTU#7 but consists of only one representative sequence/ unique amplicon.

Marker	# OTU	Species	Unique Amplicons	Total sequences	Seed sequences
V4	1	*L. aporus*	56	54,214	52,831
	2	*L. danicus* V4 clade 1	31	30,071	29,779
	3	*L. convexus*	14	18,246	18,115
	4	*L. hargravesii*	3	802	792
	5	*L. danicus* V4 clade 2	3	50	23
	7	*L. aporus*	2	17	14
	28	*L. hargravesii* DEL clade	2	8	5
V9	1	*L. aporus*	166	61,459	58,765
	2	*L. danicus/ hargravesii*	123	38,452	35,387
	3	*L. convexus*	95	18,650	18,021
	4	*T. belgicus*	10	1,775	1,676
	5	*L. aporus*	2	54	51
	6	*L. danicus/ hargravesii*	2	38	35

V4 dataset produced more OTUs but the majority (93%) consisted of only one amplicon while V9 dataset produced less OTUs but 50% consisted of more than one amplicons; therefore the number of unique amplicons per OTU was higher in V9. From the comparison of the OTUs with the phylogeny it was concluded that for both markers OTU#1, OTU#2 and OTU#3 corresponded to *L. aporus, L. danicus* and *L. convexus* sequences respectively (Fig. 5.3.1.2). In V4, OTU#4 consisted of *L. hargravesii* sequences, OTU#5 was composed by *L. danicus*_V4_clade2 sequences, OTU#7 of *L. aporus* sequences and OTU#28 was made of *L. hargravesii* DEL clade sequences.

In V9, the smaller swarm seen within the *L. aporus* OTU#2 corresponded to the *L. aporus* V9 clade2, the swarm within *L. danicus* was the *L. hargravesii* sequences, OTU#4 was *T. belgicus*, OTU#5 included a couple of *L. aporus* sequences and OTU#6 a *L. danicus* and a *L. hargravesii* sequence together due to a gap in the very beginning of the sequences. Due to the very low number of unique amplicons of these OTUs, their representation was pointless.

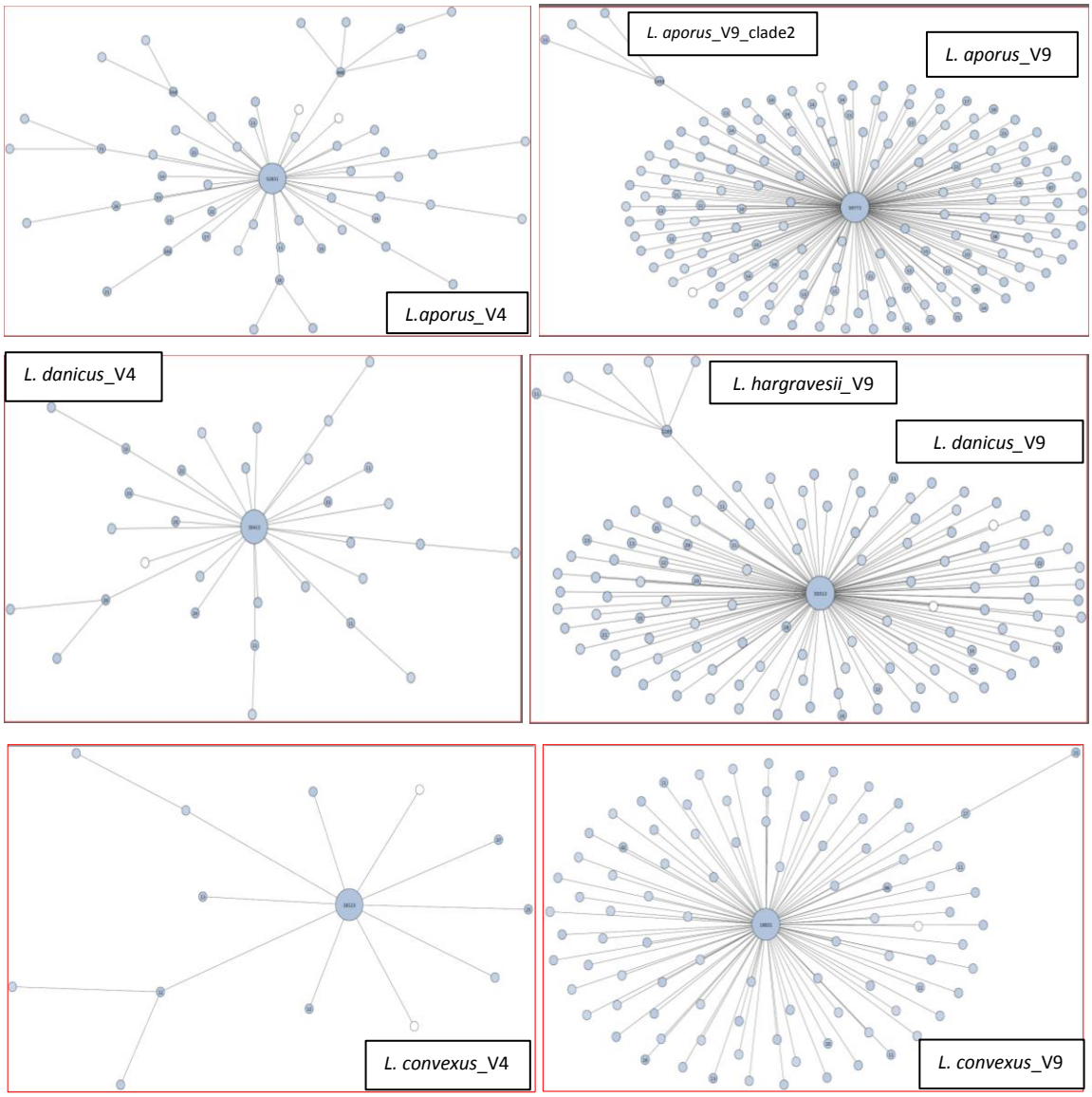


Figure 5.3.1.2 Graphs produced by Swarm for OTU#1, OTU#2 and OTU#3 in the V4 and V9 LTER-MC dataset, corresponding to *L. aporus*, *L. danicus* and *L. convexus* respectively. The central node is the seed (the size of which depends on its abundance), the representative amplicon and most abundant one for each OTU. The number within each node corresponds to the number of sequences for each amplicon (numbers lower than 10 are not shown). Each line represents a step of one difference between the two nodes.

The OTUs with a single representative sequence in V4 were mainly *L. aporus* sequences (42, following 19 of *L. danicus,* 16 of *T. belgicus,* four of *L. hargravesii* and one of the *L. minimus* reference) while in V9 there were two *L. aporus* and two *L. convexus,* one *L. danicus* and again one

of the *L. minimus* reference.

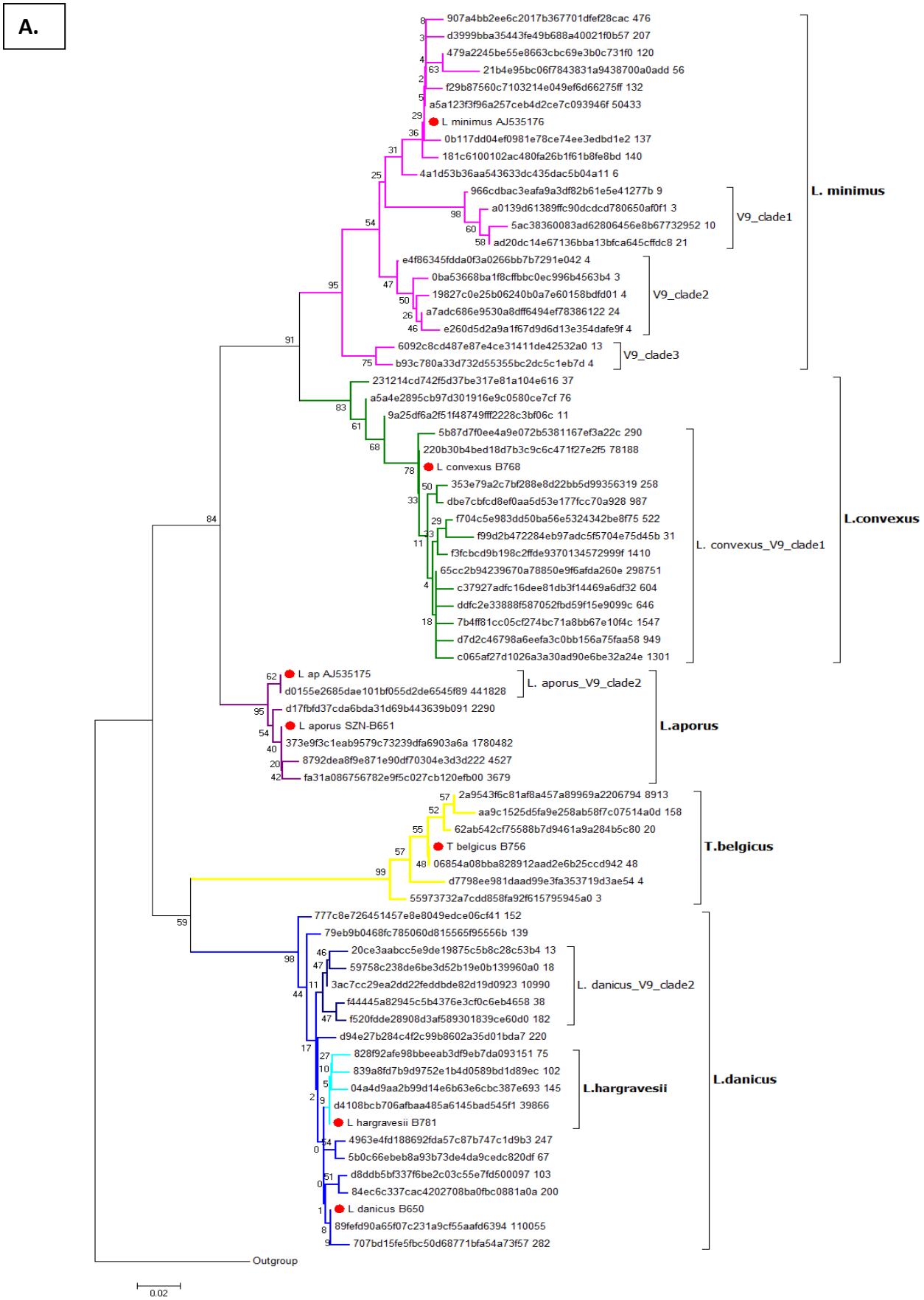
The Tara and BioMarKs V9 dataset

The local blast against Leptocylindraceae produced 3,901 sequences of which 3,030 were unique (removal of duplicates), equal or more than 90% identical to Leptocylindraceae references with length coverage equal or more than 120 bp. 2,616 ribotypes were grouped with Leptocylindraceae references and 414 ribotypes mapped alone or with other diatoms on the phylogenetic tree.

Table 5.3.1.4 Number of unique (ribotypes) and total sequences detected for each species based on V9 of the total BioMarKs and Tara dataset.

Species	N of ribotypes	N of Seqs
L. aporus	979	2,397,891
L. danicus	418	132,268
L. hargravesii	306	44,431
L. convexus	507	402,950
L. minimus	319	55,730
T. belgicus	87	9,863
Total	2,616	3,043,133

In the trees (Fig. 5.3.1.3), all species grouped separately except for *L. hargravesii* which clustered within the *L. danicus* clade. *L. aporus* V9 clade2 seen in the LTER-MC dataset was also detected here. In addition, a clade in *L. convexus* (*L. convexus* V9 clade 2) and another one in *L. danicus* (*L. danicus* V9 clade 2) were detected, which were diversified because of a single nucleotide position change. *L. minimus* was the most variable species with three subclades. As in the MC dataset, NJ and ML trees were in accordance regarding the assignment of sequences to species and clades.


A.
907a4bb2ee6c2017b367701dfef28cac 476
d3999bba35443fe49b688a40021f0b57 207
479a2245be55e8663cbc69e3b0c731f0 120
21b4e95bc06f7843831a9438700a0add 56
f29b87560c7103214e049ef6d66275ff 132
a5a123f3f96a257ceb4d2ce7c093946f 50433
L minimus AJ535176
0b117dd04ef0981e78ce74ee3edbd1e2 137
181c6100102ac480fa26b1f61b8fe8bd 140
4a1d53b36aa543633dc435dac5b04a11 6
966cdbac3eafa9a3df82b61e5e41277b 9
a0139d61389ffc90dcdcd780650af0f1 3
5ac38360083ad62806456e8b67732952 10
ad20dc14e67136bba13bfca645cffdc8 21
e4f86345fdda0f3a0266bb7b7291e042 4
0ba53668ba1f8cffbbc0ec996b4563b4 3
19827c0e25b06240b0a7e60158bdfd01 4
a7adc686e9530a8dff6494ef78386122 24
e260d5d2a9a1f67d9d6d13e354dafe9f 4
6092c8cd487e87e4ce31411de42532a0 13
b93c780a33d732d55355bc2dc5c1eb7d 4
V9_clade1
V9_clade2
V9_clade3
L. minimus
231214cd742f5d37be317e81a104e616 37
a5a4e2895cb97d301916e9c0580ce7cf 76
9a25df6a2f51f48749fff2228c3bf06c 11
5b87d7f0ee4a9e072b5381167ef3a22c 290
220b30b4bed18d7b3c9c6c471f27e2f5 78188
L convexus B768
353e79a2c7bf288e8d22bb5d99356319 258
dbe7cbfcd8ef0aa5d53e177fcc70a928 987
f704c5e983dd50ba56e5324342be8f75 522
f99d2b472284eb97adc5f5704e75d45b 31
f3fcbcd9b198c2ffde9370134572999f 1410
65cc2b94239670a78850e9f6afda260e 298751
c37927adfc16dee81db3f14469a6df32 604
ddfc2e33888f587052fbd59f15e9099c 646
7b4ff81cc05cf274bc71a8bb67e10f4c 1547
d7d2c46798a6eefa3c0bb156a75faa58 949
c065af27d1026a3a30ad90e6be32a24e 1301
L. convexus_V9_clade1
L.convexus
L ap AJ535175
d0155e2685dae101bf055d2de6545f89 441828
d17fbfd37cda6bda31d69b443639b091 2290
L aporus SZN-B651
373e9f3c1eab9579c73239dfa6903a6a 1780482
8792dea8f9e871e90df70304e3d3d222 4527
fa31a086756782e9f5c027cb120efb00 3679
L. aporus_V9_clade2
L.aporus
2a9543f6c81af8a457a89969a2206794 8913
aa9c1525d5fa9e258ab58f7c07514a0d 158
62ab542cf75588b7d9461a9a284b5c80 20
T belgicus B756
06854a08bba828912aad2e6b25ccd942 48
d7798ee981daad99e3fa353719d3ae54 4
55973732a7cdd858fa92f615795945a0 3
T.belgicus
777c8e726451457e8e8049edce06cf41 152
79eb9b0468fc785060d815565f95556b 139
20ce3aabcc5e9de19875c5b8c28c53b4 13
59758c238de6be3d52b19e0b139960a0 18
3ac7cc29ea2dd22feddbde82d19d0923 10990
f44445a82945c5b4376e3cf0c6eb4658 38
f520fdde28908d3af589301839ce60d0 182
L. danicus_V9_clade2
d94e27b284c4f2c99b8602a35d01bda7 220
828f92afe98bbeeab3df9eb7da093151 75
839a8fd7b9d9752e1b4d0589bd1d89ec 102
04a4d9aa2b99d14e6b63e6cbc387e693 145
d4108bcb706afbaa485a6145bad545f1 39866
L hargravesii B781
L.hargravesii
4963e4fd188692fda57c87b747c1d9b3 247
5b0c66ebeb8a93b73de4da9cedc820df 67
d8ddb5bf337f6be2c03c55e7fd500097 103
84ec6c337cac4202708ba0fbc0881a0a 200
L danicus B650
89fefd90a65f07c231a9cf55aafd6394 110055
707bd15fe5fbc50d68771bfa54a73f57 282
L.danicus
Outgroup
0.02

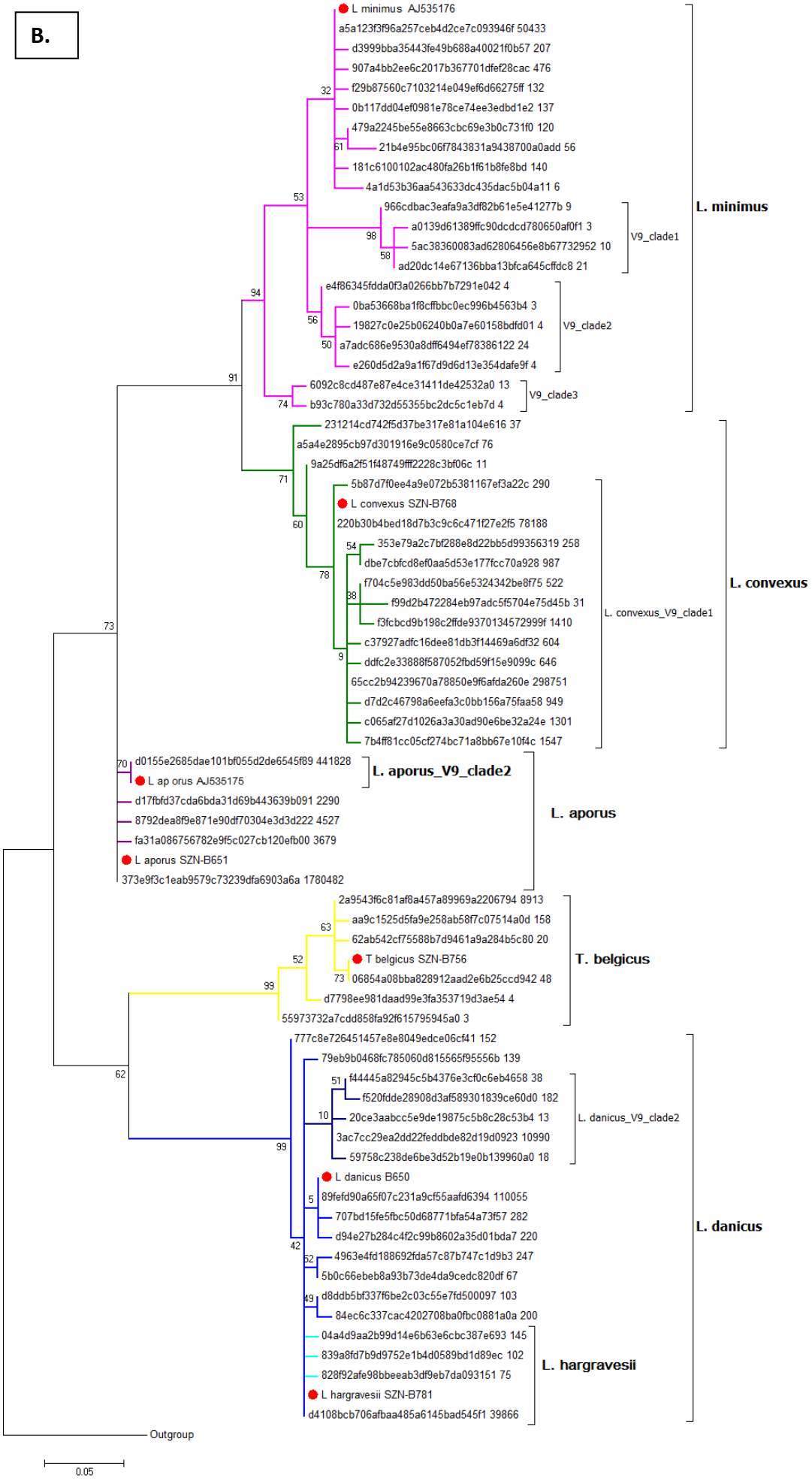

B.
L minimus AJ535176
a5a123f3f96a257ceb4d2ce7c093946f 50433
d3999bba35443fe49b688a40021f0b57 207
907a4bb2ee6c2017b367701dfef28cac 476
f29b87560c7103214e049ef6d66275ff 132
0b117dd04ef0981e78ce74ee3edbd1e2 137
479a2245be55e8663cbc69e3b0c731f0 120
21b4e95bc06f7843831a9438700a0add 56
181c6100102ac480fa26b1f61b8fe8bd 140
4a1d53b36aa543633dc435dac5b04a11 6
966cdbac3eafa9a3df82b61e5e41277b 9
a0139d61389ffc90dcdcd780650af0f1 3
5ac38360083ad62806456e8b67732952 10
ad20dc14e67136bba13bfca645cffdc8 21
V9_clade1
e4f86345fdda0f3a0266bb7b7291e042 4
0ba53668ba1f8cffbbc0ec996b4563b4 3
19827c0e25b06240b0a7e60158bdfd01 4
a7adc686e9530a8dff6494ef78386122 24
e260d5d2a9a1f67d9d6d13e354dafe9f 4
V9_clade2
6092c8cd487e87e4ce31411de42532a0 13
b93c780a33d732d55355bc2dc5c1eb7d 4
V9_clade3
L. minimus
231214cd742f5d37be317e81a104e616 37
a5a4e2895cb97d301916e9c0580ce7cf 76
9a25df6a2f51f48749fff2228c3bf06c 11
5b87d7f0ee4a9e072b5381167ef3a22c 290
L convexus SZN-B768
220b30b4bed18d7b3c9c6c471f27e2f5 78188
353e79a2c7bf288e8d22bb5d99356319 258
dbe7cbfcd8ef0aa5d53e177fcc70a928 987
f704c5e983dd50ba56e5324342be8f75 522
f99d2b472284eb97adc5f5704e75d45b 31
f3fcbcd9b198c2ffde93701345729991 1410
c37927adfc16dee81db3f14469a6df32 604
ddfc2e33888f587052fbd59f15e9099c 646
65cc2b94239670a78850e9f6afda260e 298751
d7d2c46798a6eefa3c0bb156a75faa58 949
c065af27d1026a3a30ad90e6be32a24e 1301
7b4ff81cc05cf274bc71a8bb67e10f4c 1547
L. convexus_V9_clade1
L. convexus
d0155e2685dae101bf055d2de6545f89 441828
L ap orus AJ535175
L. aporus_V9_clade2
d17fbfd37cda6bda31d69b443639b091 2290
8792dea8f9e871e90df70304e3d3d222 4527
fa31a086756782e9f5c027cb120efb00 3679
L aporus SZN-B651
373e9f3c1eab9579c73239dfa6903a6a 1780482
L. aporus
2a9543f6c81af8a457a89969a2206794 8913
aa9c1525d5fa9e258ab58f7c07514a0d 158
62ab542cf75588b7d9461a9a284b5c80 20
T belgicus SZN-B756
06854a08bba828912aad2e6b25ccd942 48
d7798ee981daad99e3fa353719d3ae54 4
55973732a7cdd858fa92f615795945a0 3
T. belgicus
777c8e726451457e8e8049edce06cf41 152
79eb9b0468fc785060d815565f95556b 139
f44445a82945c5b4376e3cf0c6eb4658 38
f520fdde28908d3af589301839ce60d0 182
20ce3aabcc5e9de19875c5b8c28c53b4 13
3ac7cc29ea2dd22feddbde82d19d0923 10990
59758c238de6be3d52b19e0b139960a0 18
L. danicus_V9_clade2
L danicus B650
89fefd90a65f07c231a9cf55aafd6394 110055
707bd15fe5fbc50d68771bfa54a73f57 282
d94e27b284c4f2c99b8602a35d01bda7 220
4963e4fd188692fda57c87b747c1d9b3 247
5b0c66ebeb8a93b73de4da9cedc820df 67
d8ddb5bf337f6be2c03c55e7fd500097 103
84ec6c337cac4202708ba0fbc0881a0a 200
L. danicus
04a4d9aa2b99d14e6b63e6cbc387e693 145
839a8fd7b9d9752e1b4d0589bd1d89ec 102
828f92afe98bbeeab3df9eb7da093151 75
L hargravesii SZN-B781
d4108bcb706afbaa485a6145bad545f1 39866
L. hargravesii
Outgroup
0.05

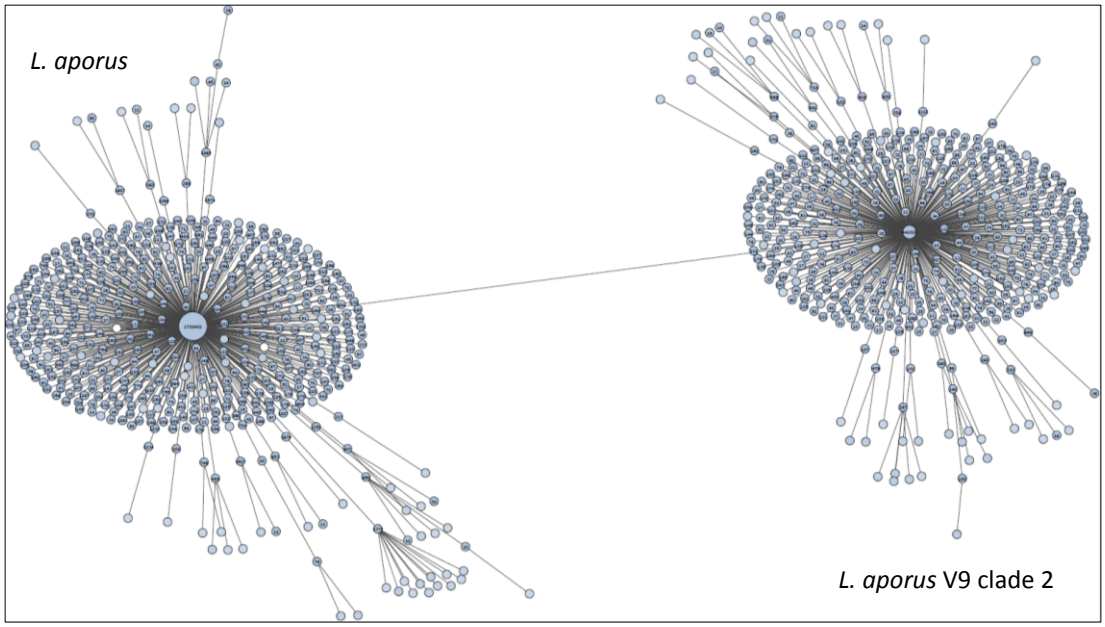
Figure 5.3.1.3 Neighbor-joining (A) and maximum-likelihood (B) tree based on V9 Leptocylindraceae sequences of all 154 stations (total BioMarKs and Tara dataset), with bootstrap method set as test of phylogeny (500 replications) and Kimura 2-parameter model as substitution model. For a clear representation of the tree here and only for this, the ten most abundant ribotypes of each species were selected. The last number in the ribotype labels represents the number of sequences.

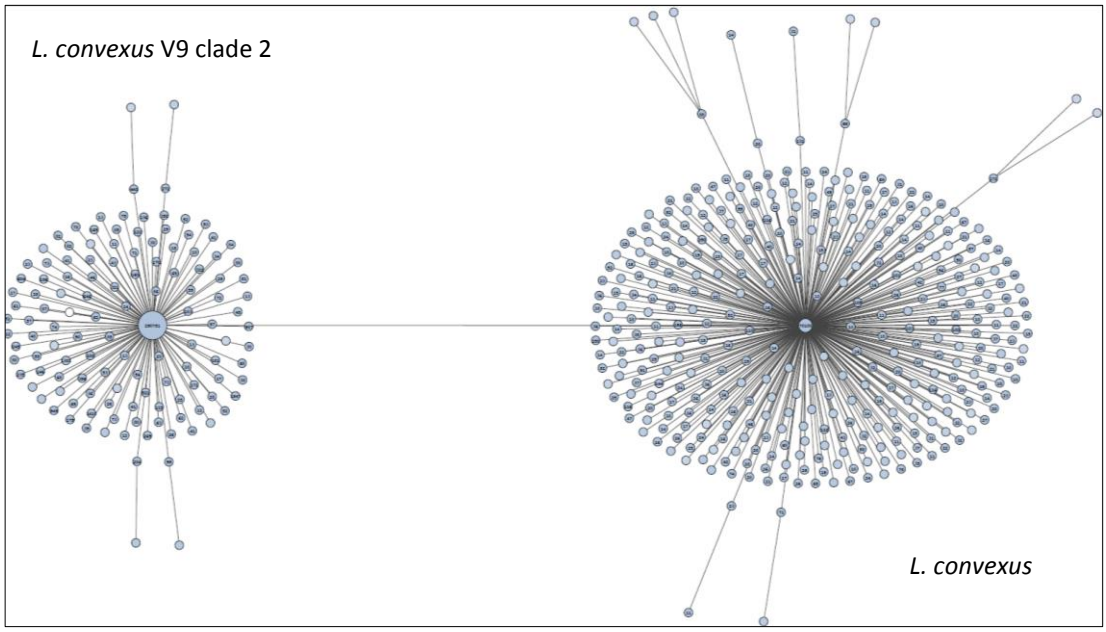
Swarm produced 29 OTUs, of which 19 consisted of only one representative sequence. The following table presents some more detailed statistics on the ten OTUs consisting of more than one representative sequence:

Table 5.3.1.5 Statistics on each OTU for total BioMarKs and Tara V9 Leptocylindraceae dataset. OTUs with only one ribotype are not presented. The seed is the representative sequence of each OTU which is also the most abundant one. The numbering of the OTUs depends on their abundance; so OTU#6 was more abundant than OTU#9 but consisted of only one representative sequence/ unique amplicon.

Marker	# OTU	Species	Unique Amplicons	Total sequences	Seed sequences
V9	1	*L. aporus*	979	2,397,443	1,780,482
	2	*L. convexus*	502	402,753	298,751
	3	*L. danicus*	724	176,592	110,055
	4	*L. minimus*	301	55,463	50,433
	5	*T. belgicus*	86	9,853	8,913
	9	*L. minimus*	2	82	75
	10	*L. convexus*	2	77	69
	13	*L. minimus*	5	39	24
	14	*L. minimus*	3	34	21

The swarm analysis confirmed the presence of the populations mentioned in the phylogenetic tree analysis (Fig. 5.3.1.4). OTU#1, OTU#2, OTU#3, OTU#4 and OTU#5 corresponded to *L. aporus, L. convexus, L. danicus, L. minimus* and *T. belgicus* respectively. The *L. aporus* OTU clearly consists of two swarms, the seed sequence of each of them being identical to the two available references, respectively. The smallest swarm corresponded to the *L. aporus* V9 clade 2, also found in the LTER-MC dataset. The same pattern was obvious in the *L. convexus* OTU, where the largest swarm corresponded to the known reference whereas the smaller one, but most abundant, differed by a single nucleotide. *L. danicus* OTU was even more complex since it consisted of three swarms. The largest one corresponded to the known *L. danicus* reference, the second to *L. hargravesii* and the smallest one to the *L. danicus* V9 clade2.


L. aporus
L. aporus V9 clade 2


L. convexus V9 clade 2
L. convexus

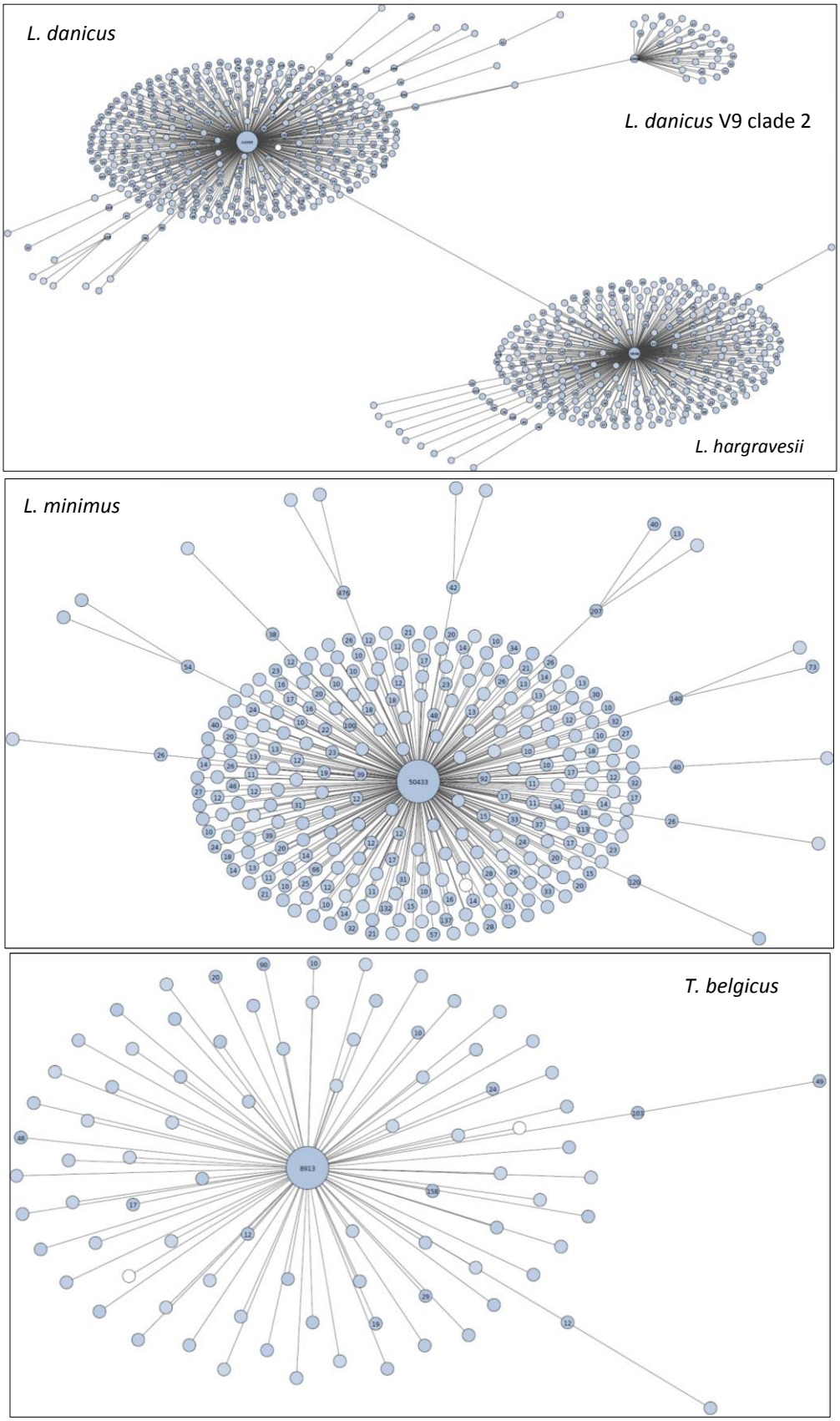


Figure 5.3.1.4 Graphs produced by Swarm for OTU#1, OTU#2, OTU#3, OTU#4 and OTU#5 in total BioMarKs and Tara Leptocylindraceae V9 dataset, corresponding to *L. aporus*, *L. convexus*, *L. danicus/ L. hargravesii*, *L. minimus* and *T. belgicus* respectively. The central node is the seed (the size of which depends on its abundance), the representative amplicon and most abundant one for each OTU. The number within each node corresponds to the number of sequences for each amplicon (numbers lower than 10 are not shown). Each line represents a step of one difference between the two nodes.

Three out of the four remaining OTUs present in the table correspond to *L. minimus* sequences while the single-sequence OTUs were eight *L. minimus,* four *L. convexus,* three *T. belgicus,* two *L. aporus* and two *L. hargravesii*, confirming what had already been noted in the phylogenetic tree about *L. minimus'* high diversity.

Summing up the diversity analysis in Leptocylindraceae of BioMarKs and Tara, the total number of species and main clades identified with V9 within species is shown in the following table:

Table 5.3.1.6 Unique (ribotypes) and total sequences of each species and main clades identified after the Leptocylindraceae diversity analysis in the BioMarKs and Tara dataset.

Species/Clade	Unique Seqs	Total Seqs
L. aporus V9 clade1	505	1,910,890
L. aporus V9 clade2	476	487,008
L. danicus V9 clade1	380	132,268
L. danicus V9 clade2	38	11,352
L. hargravesii	306	44,431
L. convexus	379	87,201
L. convexus V9 clade2	124	315,634
L. minimus	301	55,463
L. minimus V9 clade1	2	82
L. minimus V9 clade2	3	34
L. minimus V9 clade3	5	77
T. belgicus	**87**	**9,863**

5.3.2. Temporal distribution

LTER MC: V4, V9 and light microscopy data

The exact dates selected for HTS sampling (shown as stars) compared to the available weekly data from the light microscopy (LM) counts (Fig. 5.3.2.1) show a quite good coverage of the species cycle achieved by HTS over three years. However, there were few main peaks that were missed in some years, for example the *L. danicus* and *L. aporus* main peak on April 12th, 2011 and May 8th, 2012.

In all three years three main periods of increase could be distinguished: (i) mid-late spring (April or May), (ii) summer (June-July) and (iii) late summer-autumn (September-October). Peaks were separated with dates of very low concentrations or undetected presence. From late autumn (mid November) until April the cell abundances were quite low, generally <0.05 cells ml^{-1} and in some cases <0.1 cells ml^{-1}.

In 2012 the total Leptocylindraceae abundance was lower than the other two years, whereas total diatoms and phytoplankton showed a slightly lower abundance only in 2013. These interannual

differences could result to a higher representation of Leptocylindraceae in the environmental sample of 2013 and consequently led to the production of more Leptocylindraceae sequences in that year compared to 2012 when other diatom/ phytoplankton families might had compressed the actual Leptocylindraceae abundance during the sequencing process.

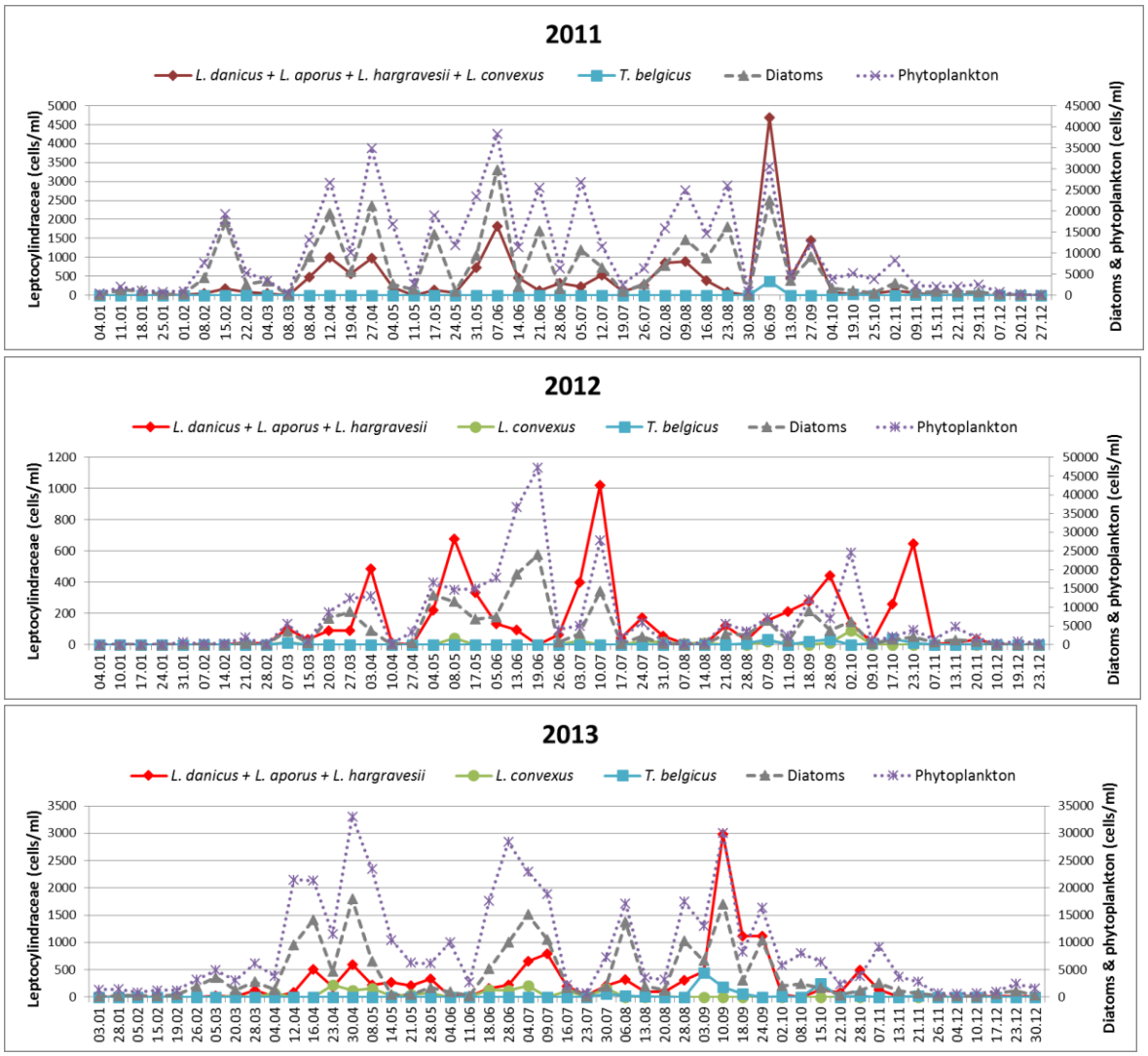
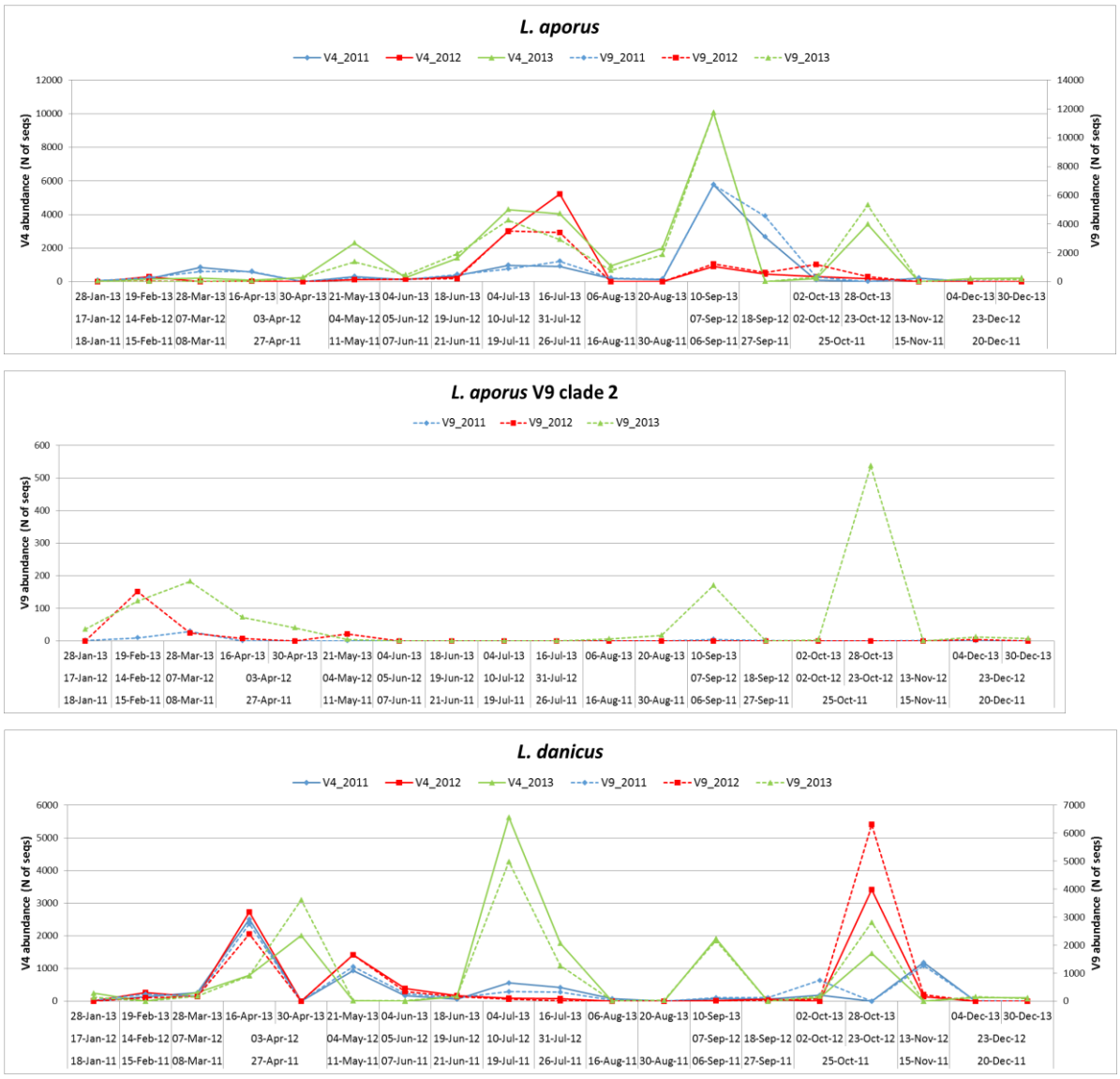


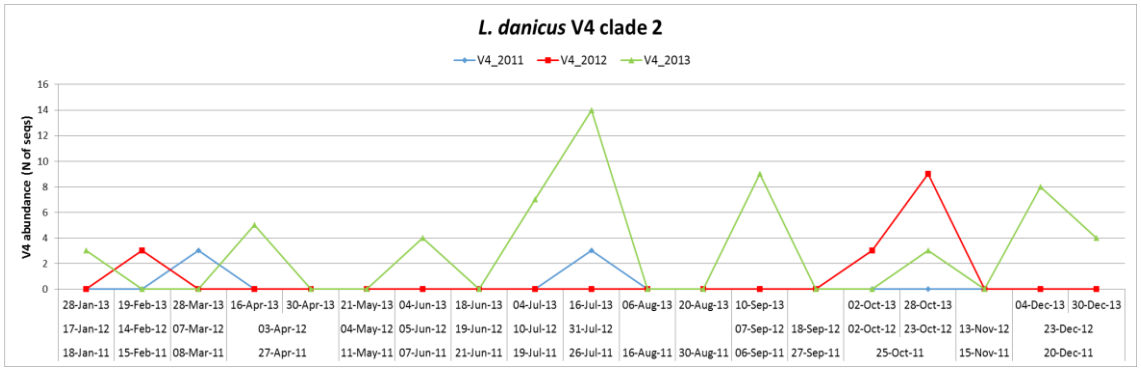
Figure 5.3.2.1 Seasonal cycles of Leptocylindraceae species, all diatoms and all phytoplankton based on light microscopy (LM) counts in 2011, 2012 and 2013 at the LTER-MC station. The sampling dates selected for HTS analysis are marked with a star. In 2011 *Leptocylindrus danicus*, *L. aporus*, *L. hargravesii* and *L. convexus* were indistinguishable. In 2012 and 2013 *L. convexus* was characterized and therefore counted separately under LM but *L. danicus* was still undistinguished from *L. aporus* and *L. hargravesii*.

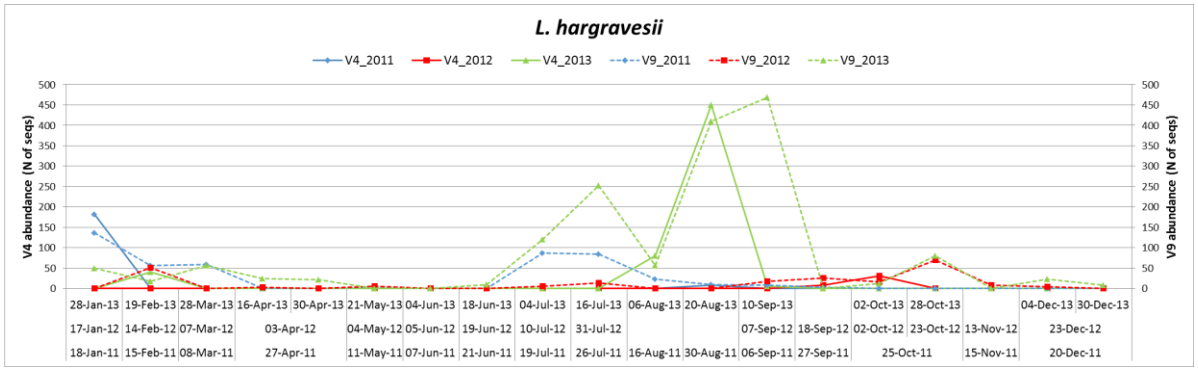
To further explore interannual variability, the average of the total Leptocylindraceae LM counts for each month was statistically tested through the three years in order to detect any possible differences in the family abundance during the time of the study. Due to non-normal distributed data and non-equal variances in most months, Welch ANOVA and/ or Kruskal-Wallis ANOVA were used. All individual months had similar abundance (non-significant difference) across the three

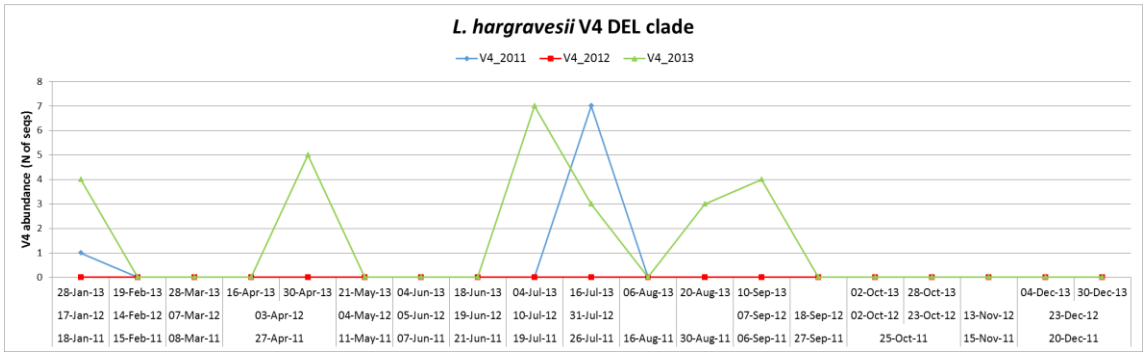
years except for February 2013 and 2011 (p=0.025) and September 2012 – 2011 (p=0.038) and 2012 – 2013 (p=0.033) which were found significantly different by the Kruskal-Wallis test. Indeed this result is also obvious in the graph (Fig. 5.3.2.1) showing that the peak corresponding to the autumn bloom in September 2012 was much lower (around 400 cells/ml) than the ones of the other two years (3,000 – 4,500 cells/ml).

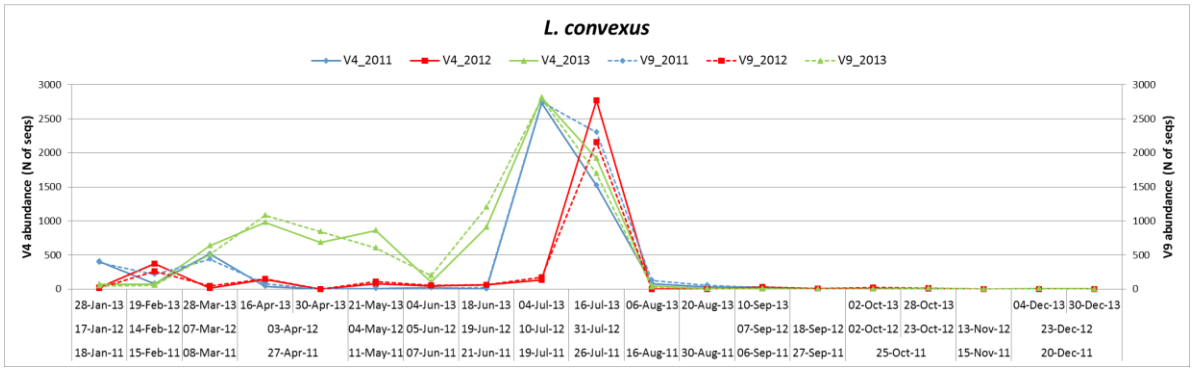
A detailed representation of the temporal distribution of each species and their related clades (section 5.3.1) based on HTS data is shown in Fig. 5.3.2.2.











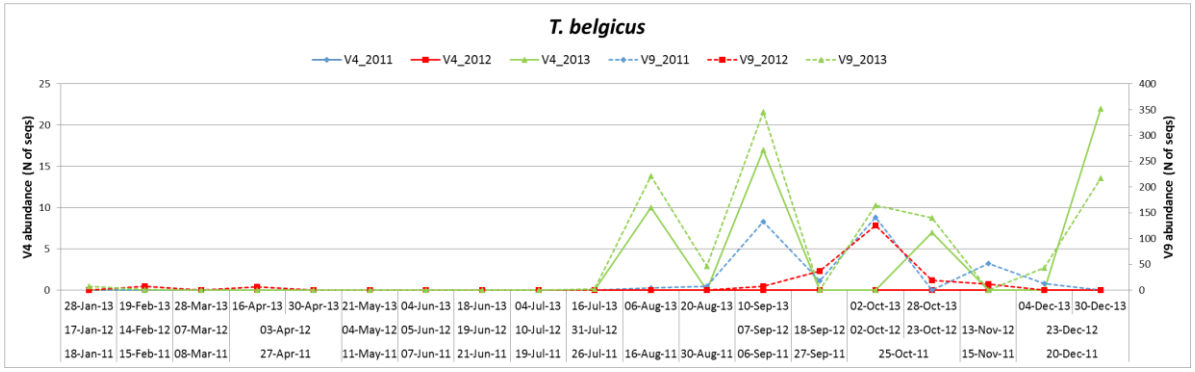


Figure 5.3.2.2 Seasonality of Leptocylindraceae species and their related clades based on the number of V4 and V9 sequences at the LTER-MC station for years 2011-2013.

For *L. aporus*, V4 and V9 sequences showed comparable seasonal distribution for the three years investigated, with peaks detected in May 2013 (2,308 V4 and 1,368 V9 sequences), July 2012 and 2013 (2,987-5,234 V4 and 2,940-4,276 V9 sequences) and September 2011 and 2013 (up to 10,059 V4 and 11,730 V9 sequences in 2013), whereas in 2013 a peak (3,416 V4 and 5,330 V9) was recorded also at the end of October. The out-grouped V4 *L. aporus* sequence seen in the phylogenetic tree (M00390_40_000000000-A6D16_1_2114_17382_11629) was represented by three individuals, all found on 27 September 2011, leading to a dead end on whether it is a real population or a product of sequencing error. *Leptocylindrus aporus* V9 clade2 was present in all three years with a significant peak of 540 sequences in October 2013, concomitant to the higher peak of the main clade. No May peak was detected in this clade whereas there was a peak in March 2013 and February 2012. It should also be mentioned that none of the *L. aporus* V4 populations could be found to correspond to the *L. aporus* V9 clade 2 since the former all had similar seasonality with no specific preference matching the one of the V9 clade 2.

Leptocylindrus danicus showed a first annual peak in April (2,005 – 3,625 sequences) of all years while in May numbers fell at ca. 1,000 in 2011 and 2012 and at 18 V4 and 8 V9 sequences in 2013. Two higher peaks were recorded in July 2013 and October 2012 and 2013 (>5,000 and 6,000 V4 or V9 sequences). Another peak of ca. 2,000 V4 and V9 sequences was recorded in September 2013. The out-grouping V4 ribotype (M00390_81_000000000-AA7DR_1_2103_9637_20702) was represented by 3 sequences detected on a single date (July 16th 2013), not allowing to clarify if it was a real population or not. *Leptocylindrus danicus* V4 clade 2 followed the same seasonal pattern as the main clade but with much fewer sequences (at most 14 sequences in July 2013).

L. hargravesii peaked at the end of August - beginning of September 2013 (ca. 450 sequences) with a smaller peak (100-250 sequences) in July 2013 and 2011. The unexpected presence of *L. hargravesii* in summer was confirmed by isolations performed in September 2014, when of 17 strains isolated, eight were *L. aporus,* six *L. hargravesii* and three *L. convexus.* The September *L. hargravesii* strains did not differ genetically from the strains isolated in other periods of the year based on ITS marker. Another small peak was present in January of 2011 and 2013 and February

of 2012 with 50 – 175 sequences, confirming the presence of the species in winter months. The *L. hargravesii* V4 DEL clade showed the highest abundance in July 2011 and 2013, with seven sequences while August – September, April and January followed with 4-5 sequences. Comparing the seasonal patterns based on V4 and V9, all dates when the *L. hargravesii* DEL clade was found in V4 were also present in the V9 pattern. So it seems that V9 cannot distinguish the two populations due to the lower variability of the V9 region.

Leptocylindrus convexus peaked in July of all three years with maxima of ca. 2,800 sequences, followed by very low sequence number or at times no detected sequences for the following months till the end of the year. In late winter-spring, smaller peaks of ca. 500 sequences were detected in February 2012 and March 2011, whereas in 2013 a higher and more dispersed peak of ca. 1,000 sequences was recorded in March, April and May.

T. belgicus was completely absent from both V4 and V9 dataset through January to July. For this species, V9 marker detected much more sequences than V4, with a peak in September-October of all three years (up to 17 V4 and 346 V9 sequences in September 2013).

Due to the considerable variability of abundance patterns among the three years, especially for certain species, the monthly average was used to get a general idea of the actual seasonal distribution of the species. A graph based on HTS numbers standardized to Leptocylindraceae monthly sum was made to see the seasonal alternation of species within the family (Fig. 5.3.2.3) and a graph based on HTS numbers standardized to each species yearly sum was made to highlight each species' seasonality (Fig. 5.3.2.4). In the first graph, the contrasting temporal distribution of the two most abundant *Leptocylindrus* species, *L. aporus* and *L. danicus*, was apparent with *L. aporus* prevailing in the period June – September and *L. danicus* dominating in April and November. *L. aporus* V9 clade 2 showed up in February – March. In January, February and March all species were present except *T. belgicus* that assumed a higher relevance in the family only in December. *L. convexus* held a considerable percentage within Leptocylindraceae from January until July and *L. hargravesii* in January and August. The standard deviation bars demonstrate high variability between the years in July for *L. danicus, L. convexus* and *L. aporus*

and in November for *L. danicus* and *L. aporus.* Indeed, the July dates of 2013 showed much more sequences of these species than the other years did while 2012 dates showed the lowest amount of sequences. November was not sampled in 2013 but 2012 samples had much less *L. danicus* and *L. aporus* sequences than 2011.

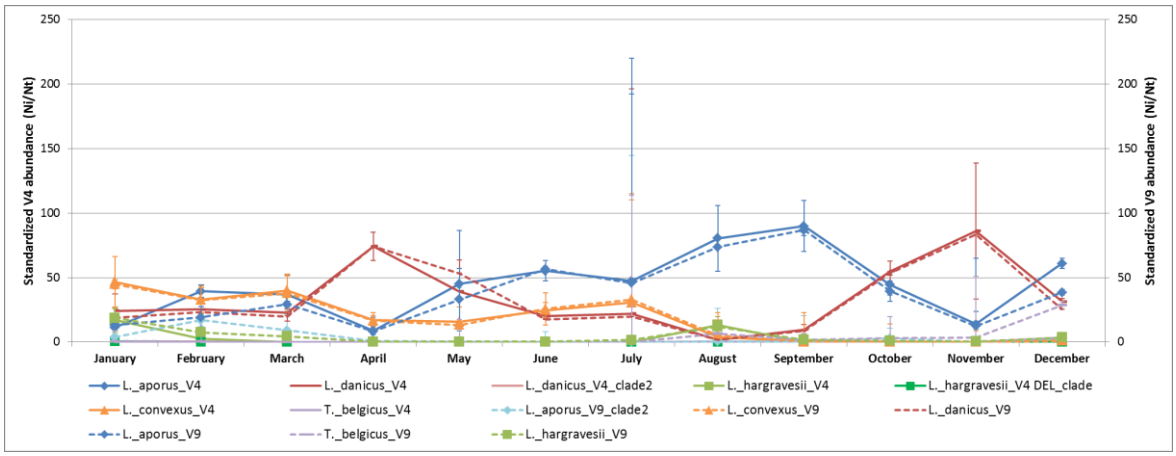


Figure 5.3.2.3 Seasonal distribution of the Leptocylindraceae species at the LTER-MC station based on the V4 and V9 average across the three years. The average abundance has been standardized to Leptocylindraceae monthly sum (Ni is the average abundance of each species/ clade and Nt is the average total abundance of all Leptocylindraceae at each month).

In the second graph (Fig. 5.3.2.4), *L. aporus* showed to be a mainly summer – autumn species with the V9 clade 2 appearing in a higher degree also in winter and spring. On the other hand, *L. danicus* was found to be a spring – autumn species with V4 clade 2 showing a higher presence in winter than the main clade. *L. hargravesii* preferred winter, autumn and summer while *L. convexus* was mainly a spring – summer species and *T. belgicus* an autumn – winter species.

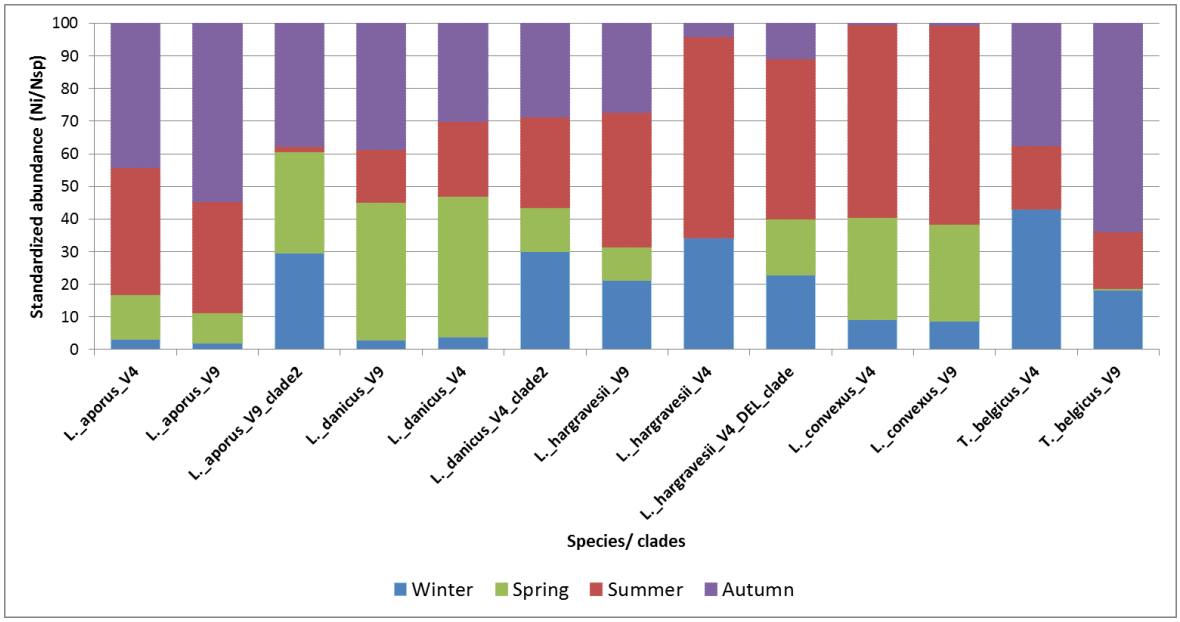
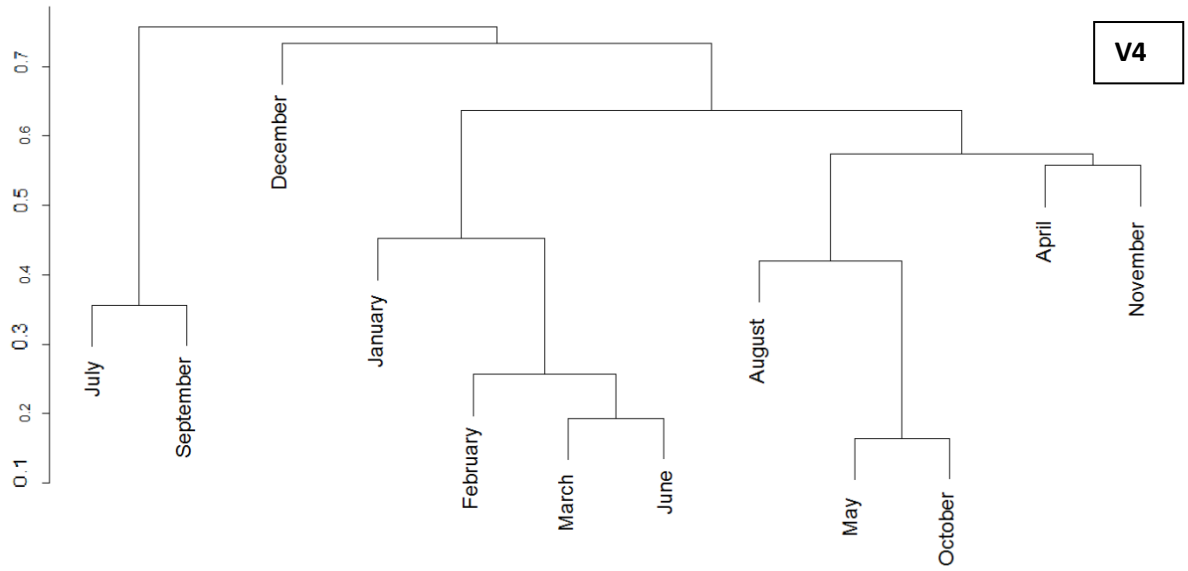


Figure 5.3.2.4 Seasonal signal for Leptocylindraceae species and clades in GoN based on the HTS V4 and V9 data. Bars indicate the proportional abundance of the sequences for each species/clade in the different seasons (Ni is the average abundance of each species/ clade and Nsp is the average total abundance of the species or clade at each year).

In addition, a hiearchical clustering and CCA analysis were performed on the average sequence abundances of each species of the three years (Fig. 5.3.2.5 and 5.3.2.6).



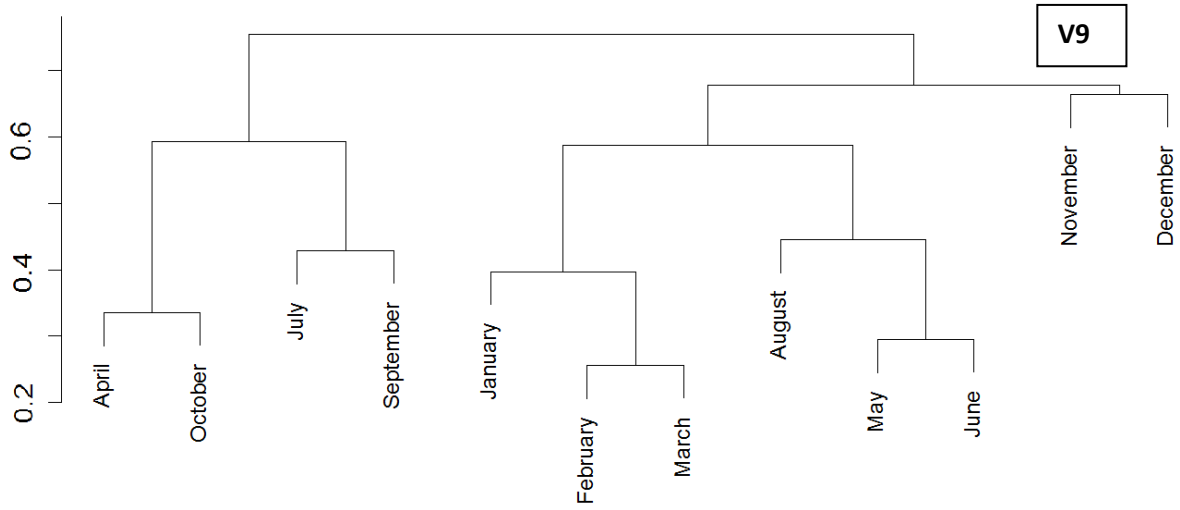
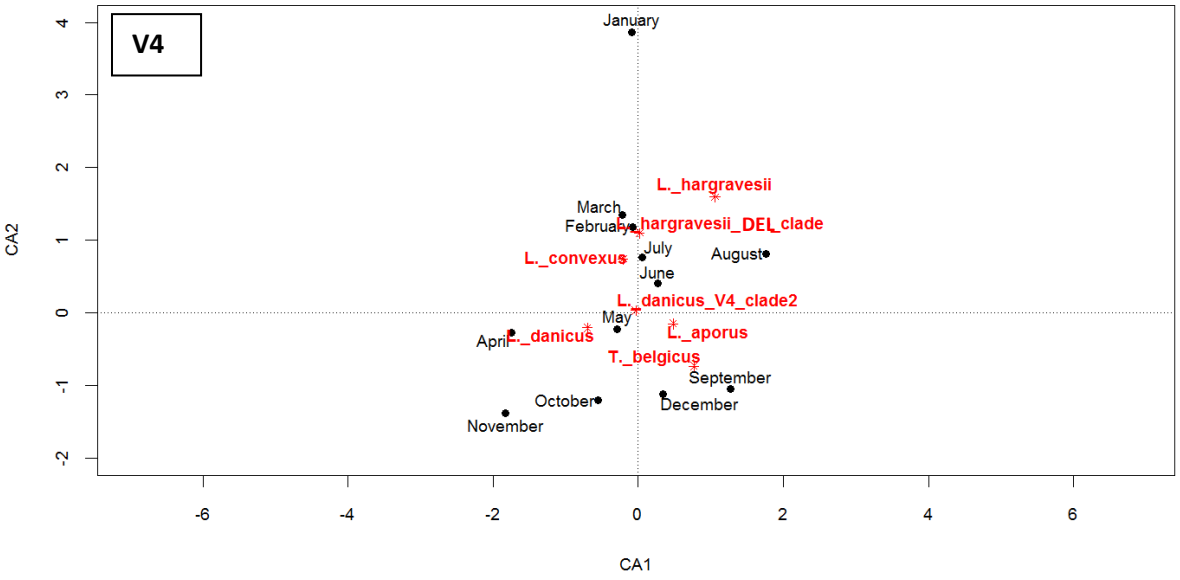


Figure 5.3.2.5 Hierarchical clustering plots on the averages of the three years of V4 and V9 abundances at the LTER-MC station.

With some exceptions, months did not cluster according to seasons but rather to the blooming of the same species in these months possibly due to similar environmental conditions. So April clustered with November in V4 or October in V9, May with October in V4 or June in V9 while January, February and March were together in both cases as well as July and September. December was the month with highest difference from the rest in V4. The slight differences between the markers' clustering result were due to their different detection power on Leptocylindraceae species.



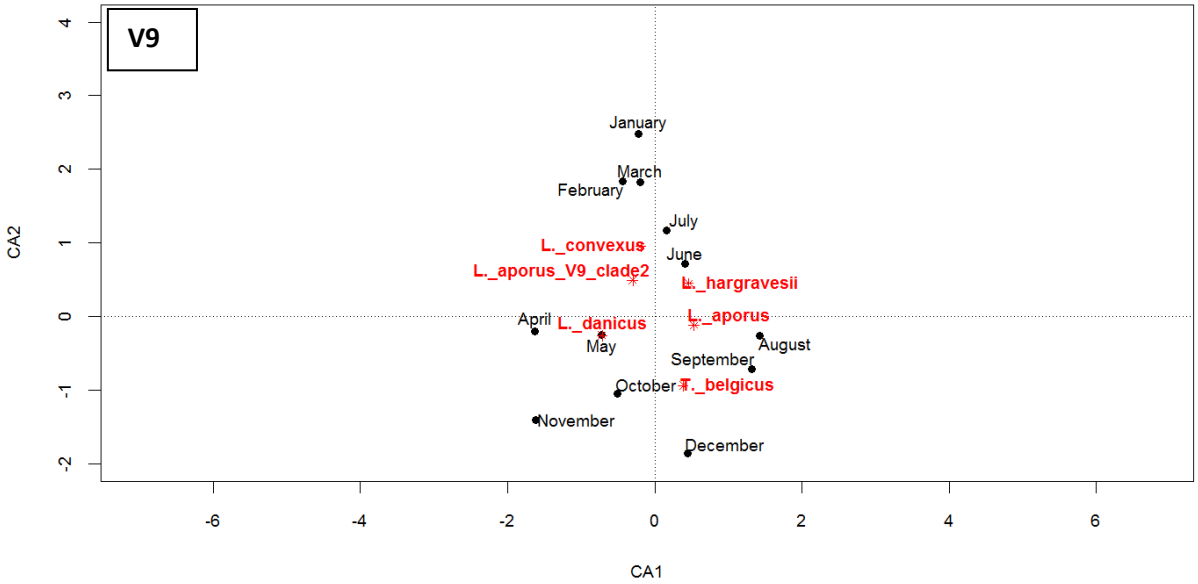


Figure 5.3.2.6 CCA analysis' plots on the averages of the three years of V4 and V9 abundances at the LTER-MC station.

The CCA plots showed the same result since there was not clear grouping of spring, autumn etc. The plots of the two markers were very similar with only few differences. January in particular was much more distant from the rest of the months in V4 plot and this seems to be a result of the *L. hargravesii* DEL clade that was undetected in V9. Other than that, April, November and October were placed slighlty away due to the *L. danicus* dominance in these months while *T. belgicus* placed December but also September away. August was characterized by *L. hargravesii* main clade and less by *L. aporus* in V4 but in V9 *L. aporus* was closer to this month.

A better discrimination between the species under the light microscope had only been achieved in 2013 since that was the year Nanjappa et al. were able to fully describe all the six different Leptocylindraceae species. Even after that, certain species were impossible to differentiate such as *L. hargravesii* and *L. danicus* while others could be difficult to tell apart when their condition was not optimal (*L. danicus* and *L. aporus)*. *L. convexus* was not distinguished under LM before 2012 and was probably ill-recognized initially in 2012 while thereafter a period of growth was identified from April through July in 2013 (Fig. 5.3.2.7).

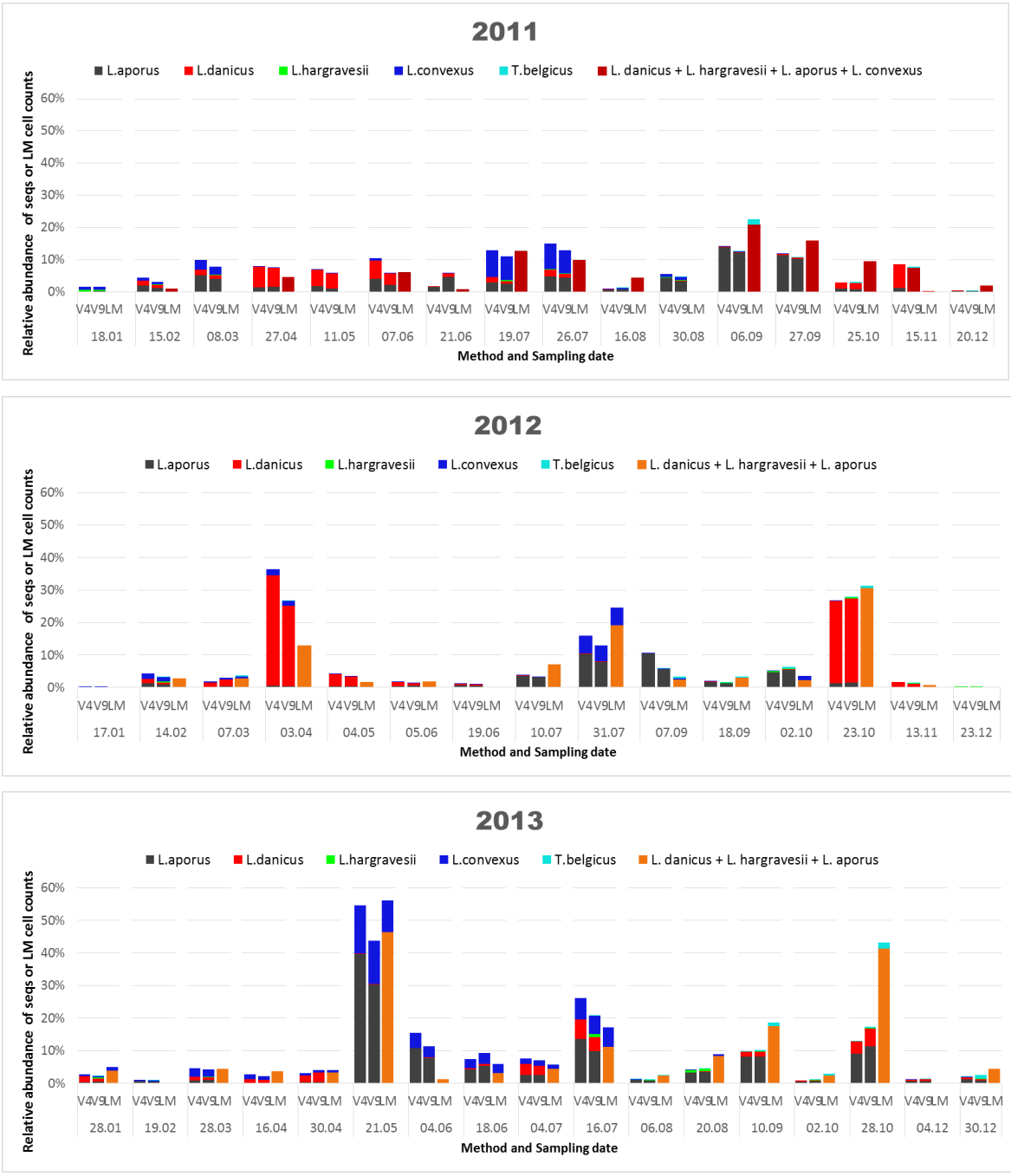
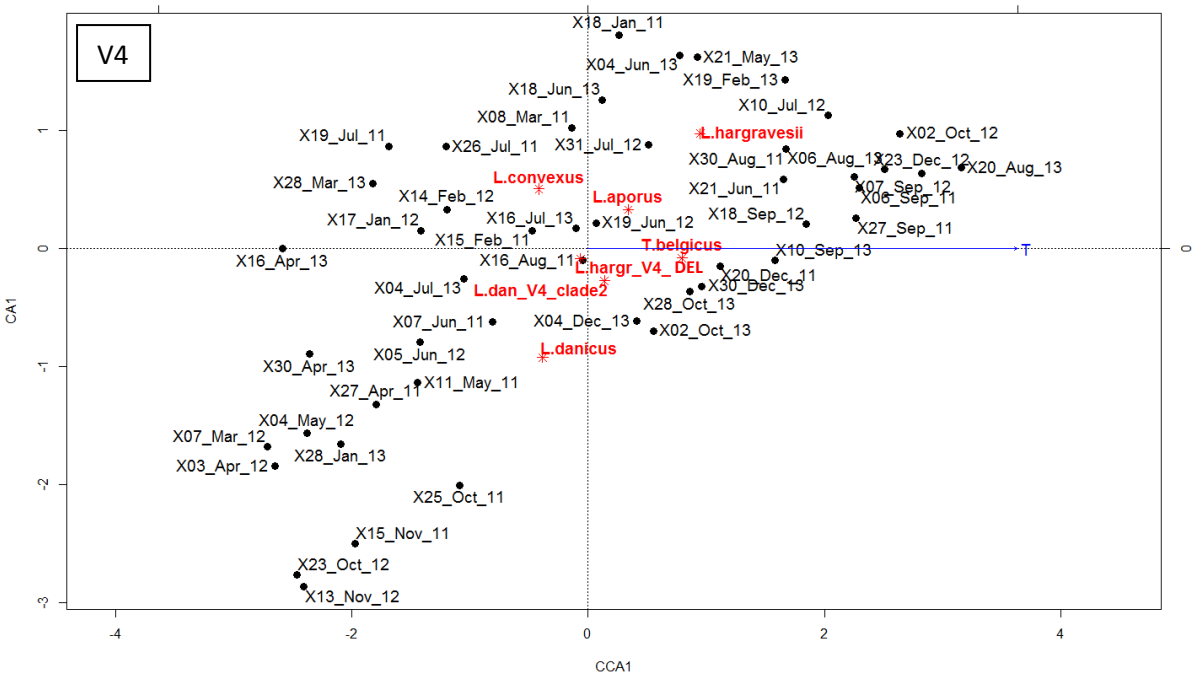


Figure 5.3.2.7 Relative abundance of HTS V4 and V9 Leptocylindraceae sequences and Leptocylindraceae cells counted under the light microscopy for 2011, 2012 and 2013 over total diatoms at LTER-MC station. In 2011 *L. danicus, L. hargravesii, L. aporus* and *L. convexus* were indistinguishable under LM (dark red). In 2012 and 2013 *L. convexus* was characterized and therefore counted separately under LM (dark blue for *L. convexus,* as in HTS, and orange for *L. danicus, L. hargravesii* and *L. aporus*).

Taking into account the limitation of the comparison for the above-mentioned reasons, in all years V4 and V9 showed high correspondence with light microscopy with only few discrepancies (April 2012 and end of October 2013) that could be due to several technical factors such as the exponential nature of the PCR technique and/or the variability of the extraction efficiency.

Environmental data

The bioenv function used on the environmental parameters calculated an environmental similarity matrix which, along with the similarity matrix of the community, was used for the calculation of a correlation coefficient. The correlation coefficient is always between 1 (perfect positive correlation) and -1 (perfect negative correlation). The best subset of environmental variables was then identified and further subjected to permutation tests to determine significance. In that way it was found that the environmental variable that best correlated with the sample similarities of the biological community (species abundance) was temperature, with a considerably low but positive Spearman correlation of 0.125 for the V4 dataset and 0.326 for the V9 dataset. A CCA analysis on the Leptocylindraceae matrix and the variables selected with bioenv showed no specific grouping of dates; the axes explained 16% of total variance in V4 and 29% in V9 dataset (Fig. 5.3.2.8). For the CCA of V9 dataset, nitrites (NO_2) were also included since temperature along with NO_2 showed a correlation coefficient of 0.3062, very close to the one of temperature alone.



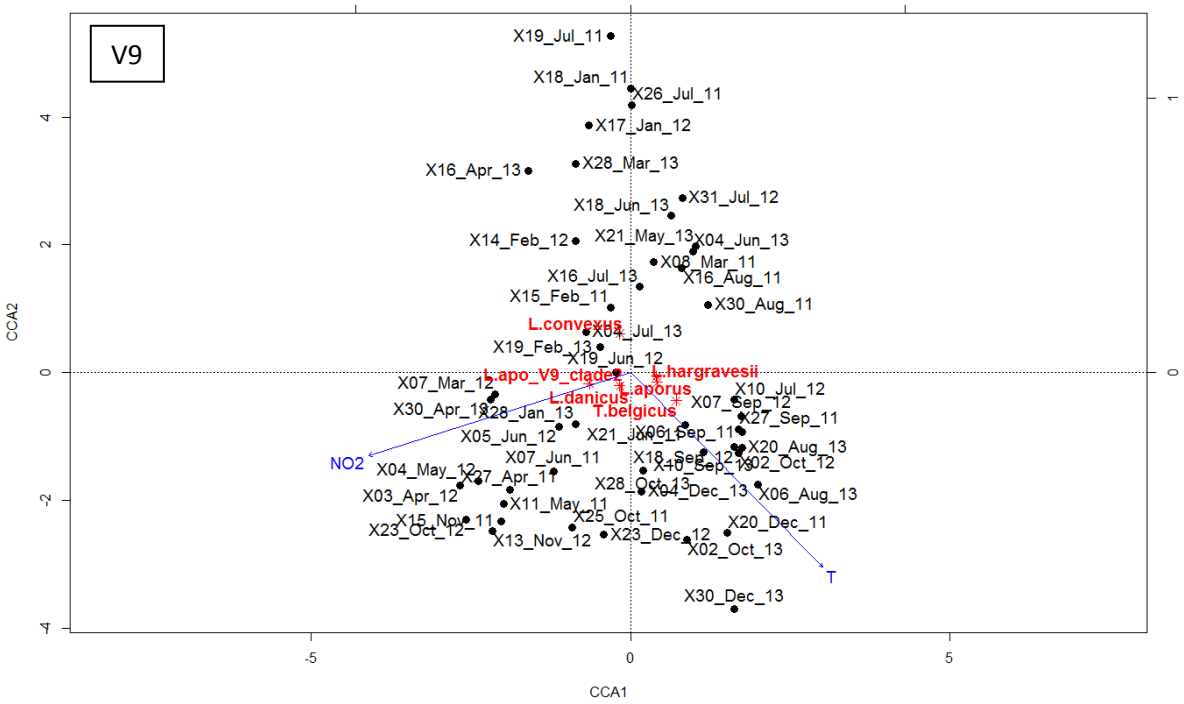


Figure 5.3.2.8 CCA plot of Leptocylindraceae V4 (above) and V9 (below) based community matrix and selected environmental parameters (temperature for V4; temperature and NO_2 for V9 dataset) for the HTS sampling dates in the three study years at the LTER-MC station. In V4 dataset, CCA1 explained 16.08% of total variance. In V9 dataset, the axes explained 28.76% (26.1% by CCA1 and 0.07% by CCA2) of total variance.

The environmental parameters of each sampling date were also used for a hierarchical clustering and CCA analysis in order to better detect any differences in the environment between the years (Fig. 5.3.2.9 and 5.3.2.10).

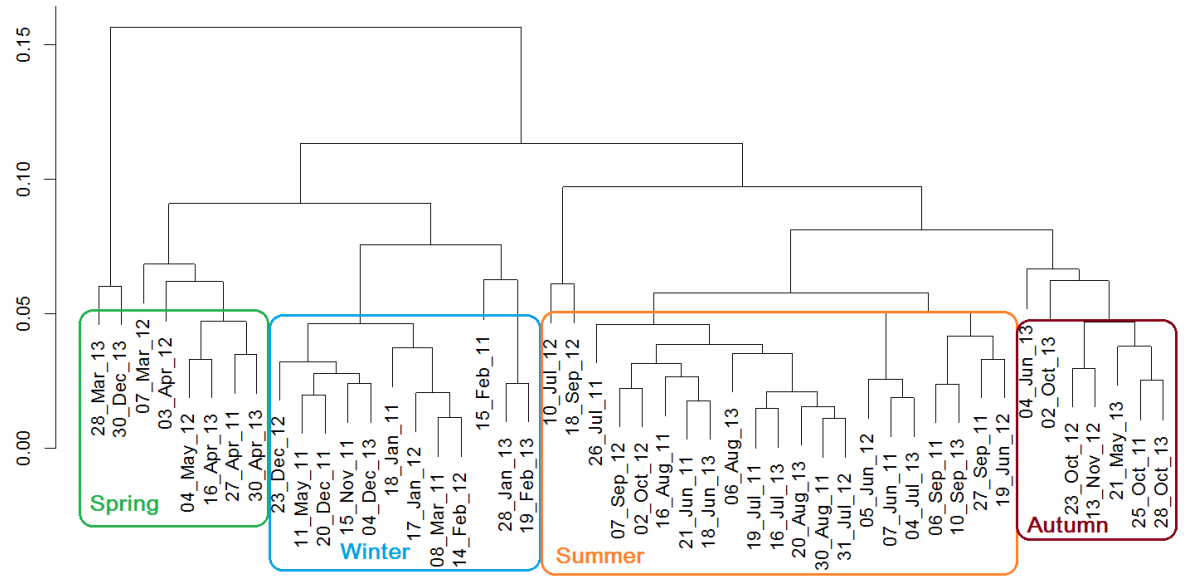


Figure 5.3.2.9 Hierarchical clustering of the environmental parameters (salinity, PO4, NH4, NO2, NO3, temperature, SiO2) for the HTS sampling dates in the three years (2011, 2012, 2013) at the LTER-MC station. The height in the clustering represents the value of the distance metric between clusters. Clades have been grouped in seasons based on the main months that constitute them.

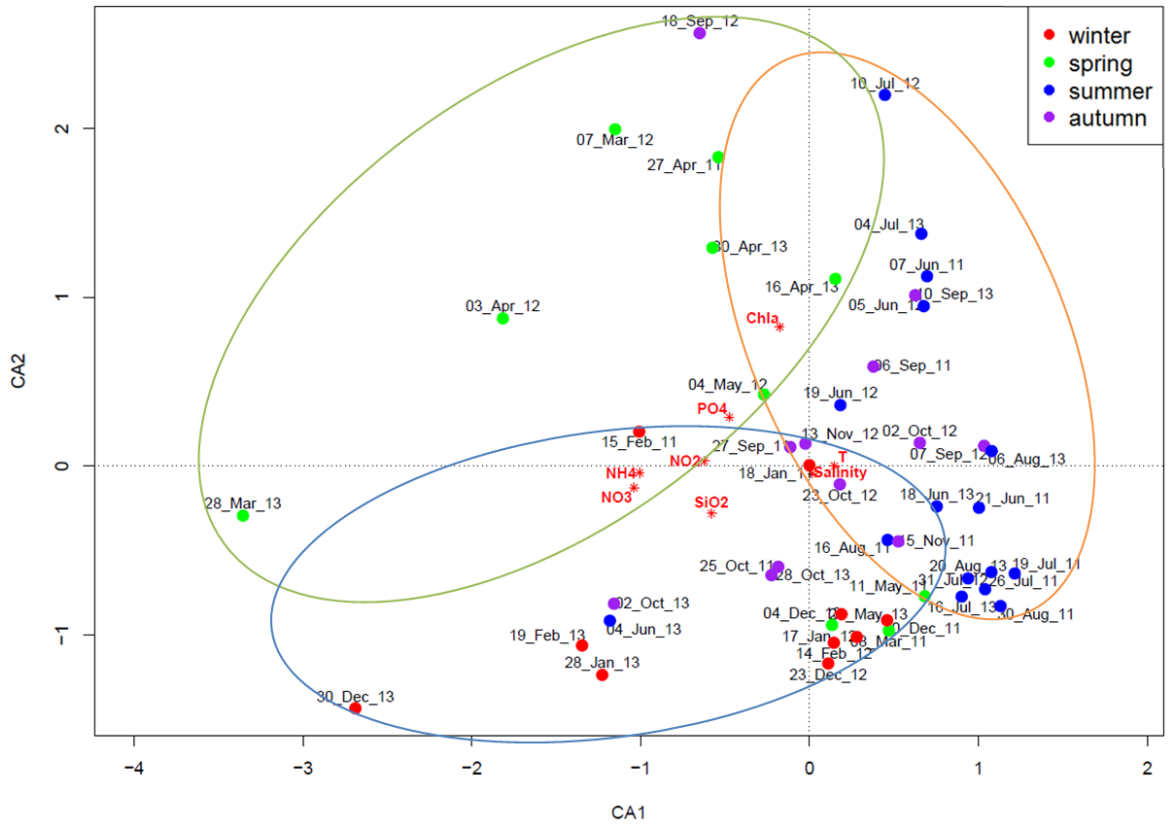


Figure 5.3.2.10 CCA plot of environmental parameters (salinity, PO4, NH4, NO2, NO3, temperature, SiO2) for the HTS sampling dates in the three study years at the LTER-MC station. The orange cycle highlights the summer-autumn dates, the green highlights the spring months and the blue cycle highlights the winter months.

Even though the hierarchical clustering and the CCA analysis showed mainly a season related clustering/grouping, there were some dates that violated this pattern. In particular, in the hierarchical clustering, the spring cluster was interrupted by a winter month, December of 2013, while on the other hand the winter cluster included two spring months, May and March of 2011. Also June and May 2013 were present in the autumn cluster while 2012 seems to be a year with the expected environmental ranges for each season. Looking at the CCA plot, September 18th 2012 was found much further than the rest of the September dates and therefore resulting in different seasonal cycles. Looking also the raw data of the environmental parameters, NH4, then NO2 or NO3 and lastly SiO2 seemed to be the ones responsible for these discrepancies. NH4 and SiO2 averages were higher in 2013. Salinity and temperature were the most stable ones within the seasons of all years. Chlorophyll a was mostly linked with the spring and summer months when the blooms of most phytoplankton species occur; its average was also higher in 2012, possibly a good year for most species.

5.3.3 Spatial distribution

Tara stations: V9 data

In the following map (Fig. 5.3.3.1), all stations explored for Leptocylindraceae presence are shown. The stations explored already by Nanjappa (2012) are included. There were only nine Tara stations where no *Leptocylindrus* species was detected, four of them overlapping in the same region of South Pacific Ocean at the map.

Figure 5.3.3.1. All stations explored for Leptocylindraceae presence in the current study. The BioMarKs station are coloured green while stations where no Leptocylindraceae was detected are red.

After the removal of the BioMarKs stations' sequences, which were only used for the biodiversity analysis, the dataset was further divided based on depth and size fraction in order to study the world distribution. A description of the number of sequences assigned to each species follows.

Table 5.3.3.1 Number of Leptocylindraceae ribotypes and the respective sequences found in the whole Tara dataset (all depths and size fractions) as well as in the surface and deep sea samples of 5 – 20 µm size fraction.

	Tara V9					
Raw Seqs	1,775,314,734					
	All size fractions and depths		**SUR, 5-20µm**		**DCM, 5-20µm**	
	N of ribotypes	**N of Seqs**	**N of ribotypes**	**N of Seqs**	**N of ribotypes**	**N of Seqs**
L. aporus	946	931,336	911	680,584	759	58,189
L. danicus	407	58,101	339	17,938	219	3,832
L. hargravesii	306	35,892	252	9,373	192	6,455
L. convexus	506	393,231	483	112,412	452	42,810
T. belgicus	24	509	12	246	7	66
L. minimus	316	36,804	266	14,511	11	139
Total Number	2,505	1,455,873	2,263	835,064	1,640	111,491

In all three sets *L. aporus* was the more abundant Leptocylindraceae species, *L. convexus* followed and then *L. danicus, L. minimus* and lastly *T. belgicus.*

The world distribution of each species and their related clades depicted on the following maps was based on the surface and DCM samples of the 5-20 µm size fraction.

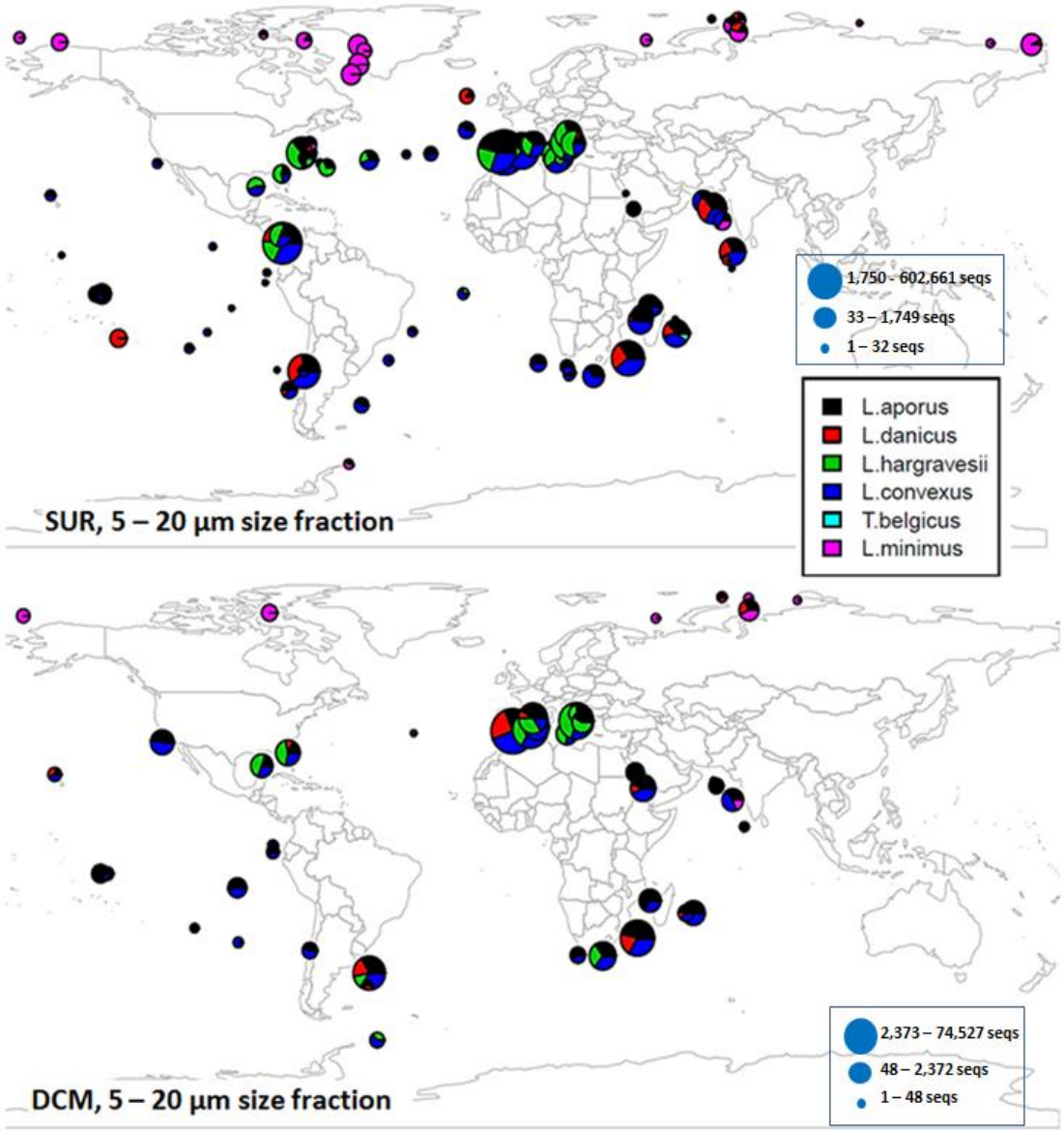


Figure 5.3.3.2 World distribution of log (abundance+1) of Leptocylindraceae based on the Tara Ocean and Tara Arctic datasets at surface and DCM, 5-20 µm size fraction.

Leptocylindraceae species were found mainly in the coastal rather the open ocean stations. The family was detected at 89 stations in surface samples and 47 stations in the DCM samples out of the 146 total Tara stations. Separate maps for each species were made for the surface samples. The separate DCM maps can be found in the appendix.

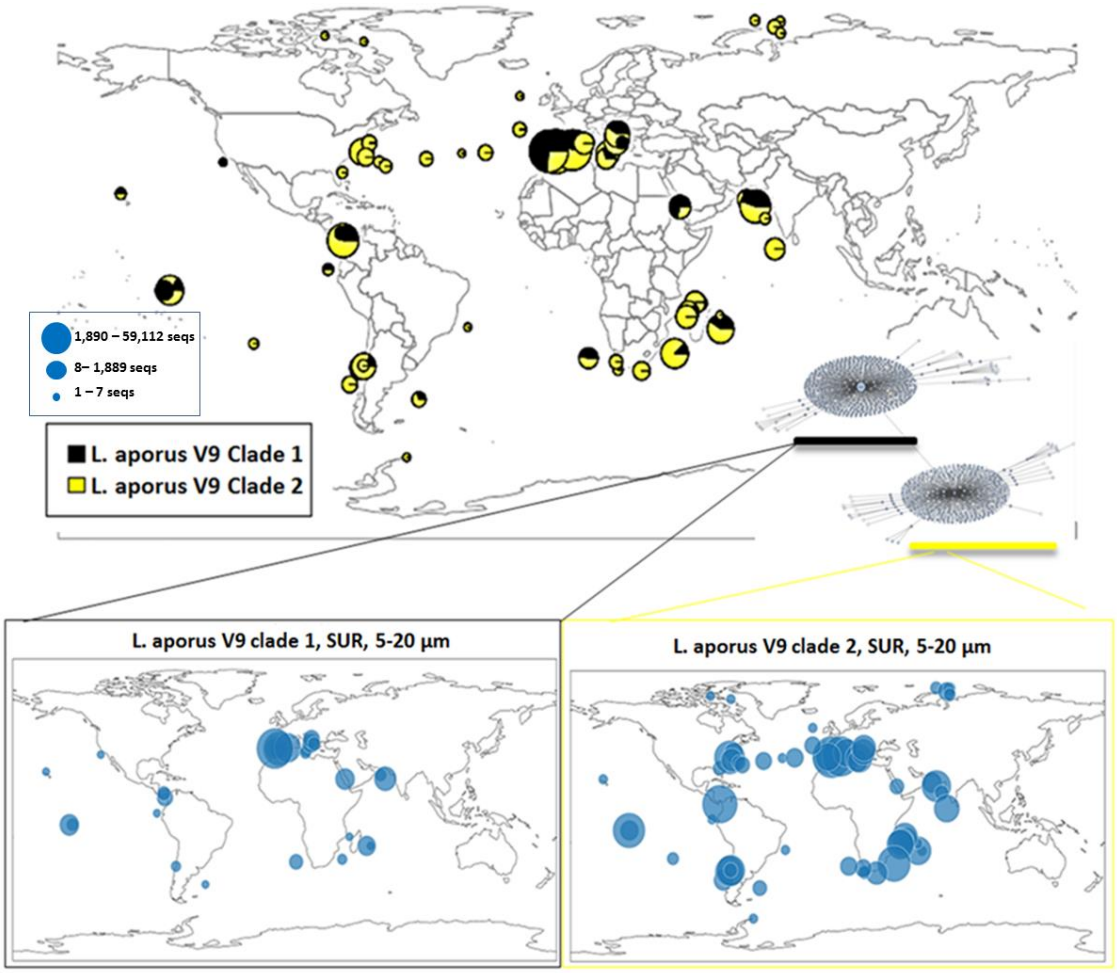


Figure 5.3.3.3 World distribution of log (abundance+1) of *L. aporus* clades at the Tara stations' surface samples, 5-20 µm size fraction. OTUs were represented by swarms (Mahè et al., 2014) and were linked to the spatial distribution with corresponding colours. The size of the bubbles in the lower maps represents the abundance within each clade.

In the surface, *L. aporus* V9 clade 1 showed some of the highest abundance numbers at certain stations while in the DCM the situation was reversed for these same stations with *L. aporus* V9 clade 2 dominating instead. Besides that, *L. aporus* V9 clade 2 was more widespread, dominating in more stations compared to V9 clade 1 in both surface and DCM.

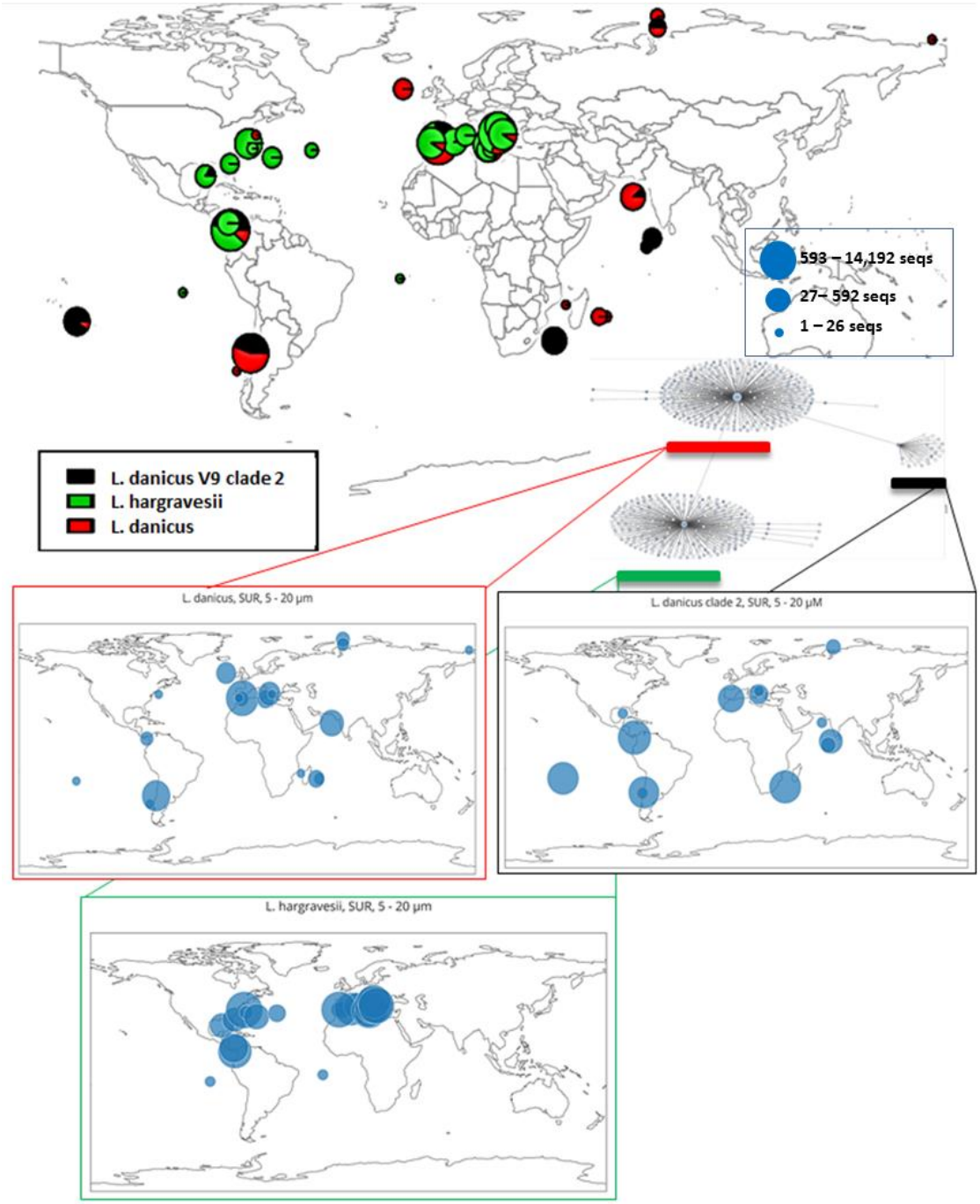


Figure 5.3.3.4 World distribution of log (abundance+1) of *L. danicus* clades and *L. hargravesii* at the Tara stations' surface samples, 5-20 µm size fraction. OTUs were represented by swarms (Mahè et al., 2014) and were linked to the spatial distribution with corresponding colours. The size of the bubbles in the lower maps represents the abundance within each clade.

In surface, *L. danicus* showed the highest abundance at certain stations while in DCM this was true for *L. hargravesii.* Besides that, *L. hargravesii* was more abundant at more stations at both depths but also had a more restricted distribution while both *L. danicus* clades are more widespread across latitudes.

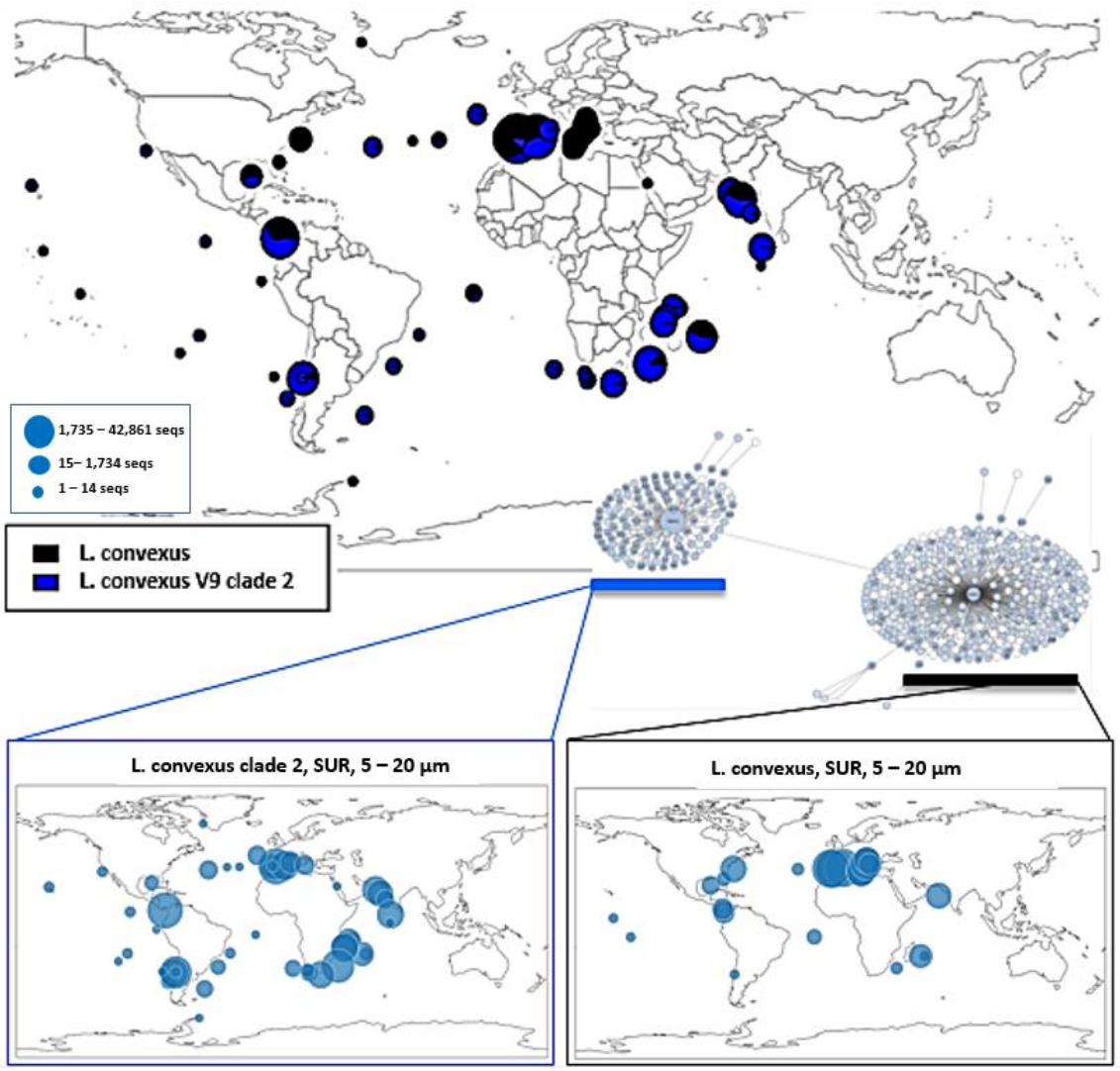


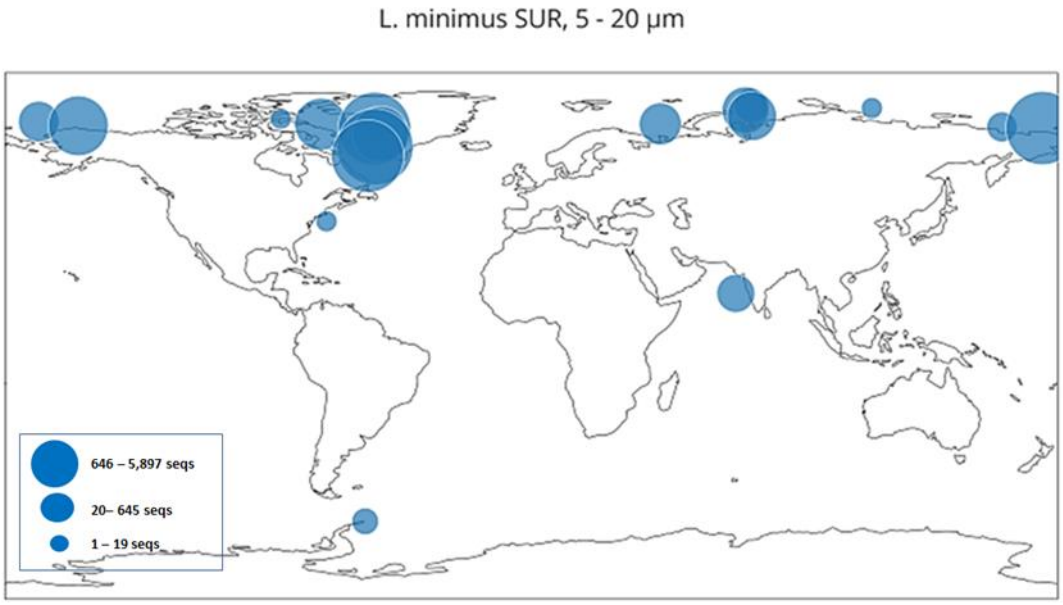
Figure 5.3.3.5 World distribution of log (abundance+1) of *L. convexus* at the Tara stations' surface samples, 5-20 µm size fraction. OTUs were represented by swarms (Mahè et al., 2014) and were linked to the spatial distribution with corresponding colours. The size of the bubbles in the lower maps represents the abundance within each clade.

Leptocylindrus convexus V9 clade 2 was more abundant but less diverse than the main clade in surface and DCM samples. Yet, the distribution of the clades was the same at both depths. A summary of the species clades' abundances depicted in the maps above is shown in Table 5.3.3.2.

Table 5.3.3.2 Number of ribotypes and total sequences for all the clades within *L. aporus, L. danicus* and *L. convexus* for surface and DCM Tara samples at 5 – 20 µm size fraction.

Species' clades	SUR, 5 – 20 µm		DCM, 5 – 20 µm	
	Ribotypes	Seqs	Ribotypes	Seqs
L. aporus V9 clade 1	472	604,230	334	11,020
L. aporus V9 clade 2	439	76,354	425	47,169
L. danicus	309	15,855	211	3,725
L. danicus V9 clade 2	30	2,081	8	107
L. hargravesii	252	9,375	192	6,455
L. convexus	362	28,854	338	13,997
L. convexus V9 clade 2	121	83,529	114	28,813

The *L. minimus* and *T. belgicus* maps show the polar distribution of the former and the highly restricted distribution of the latter (Fig.5.3.3.6). Two stations violate the *L. minimus* pattern though, in the North Atlantic and Indian Oceans. The sequences present in these stations belong to the main *L. minimus* clade and do not constitute a separate population from the Arctic one.



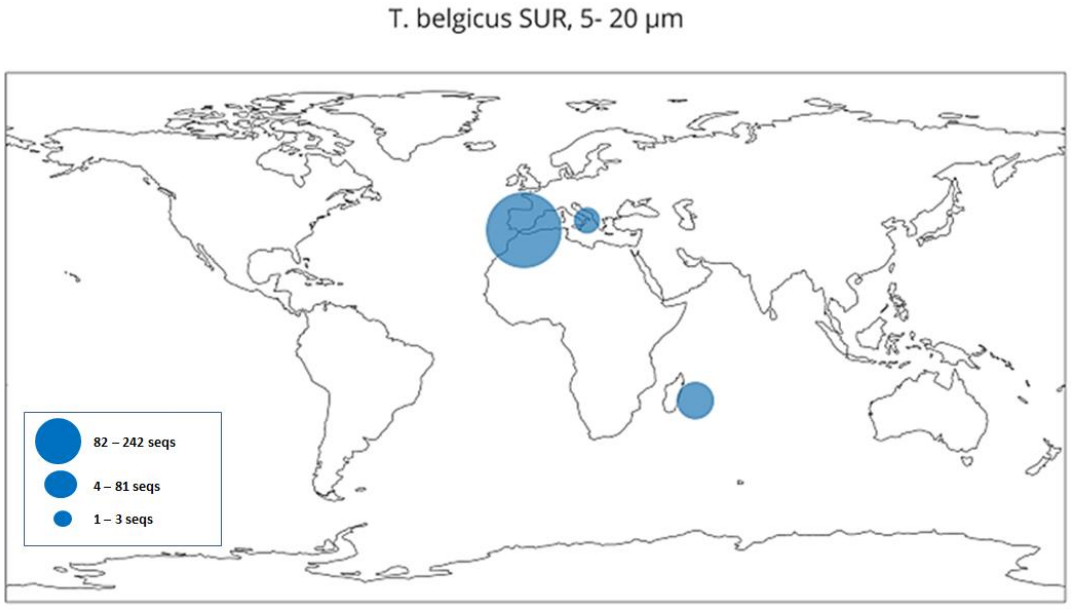
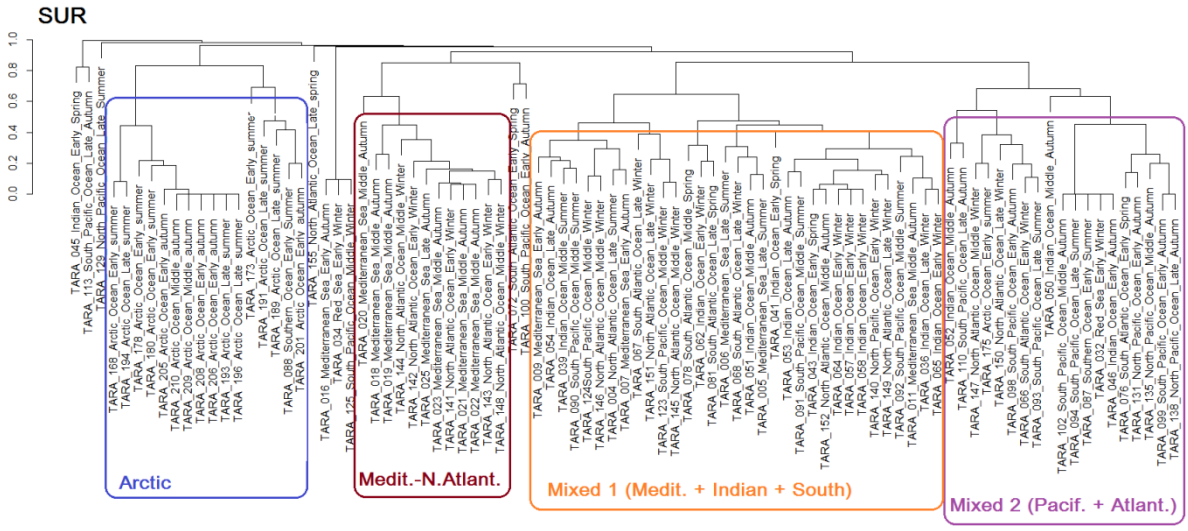


Figure 5.3.3.6 World distribution of *L. minimus* and *T. belgicus* in Tara surface samples, 5 -20 µm size fraction. The size of the bubbles represents the normalized abundance, log (abundance+1), within each clade.

The hierarchical clustering on the stations based on Leptocylindraceae rarefied abundances showed a main clustering based on geographical position but there were also signs of seasonal diversification. For example in both depths the Arctic clade consisted of stations sampled mainly in summer while the so called Mediterranean/ North Atlantic clade was also characterized by sampling in autumn. However the rest of the clades didn't show such a strong seasonal clustering.



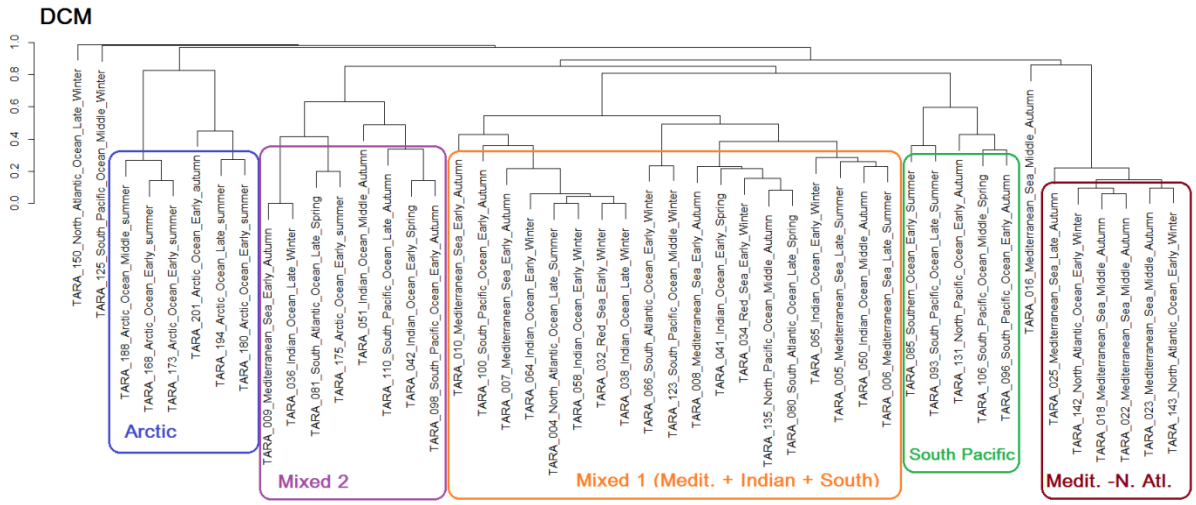


Figure 5.3.3.7 Hierarchical clustering of Tara stations based on Leptocylindraceae rarefied abundances in the 5 – 20 size fraction at surface and DCM samples.

Of course a linkage between stations and seasons cannot be avoided since the stations that were geographically near were also sampled in the same season (Fig. 5.3.3.8).

Figure 5.3.3.8 World map distribution of the Tara stations where Leptocylindraceae were detected in surface and DCM, 5-20 size fraction samples. The colour of each station is representative of the sampling season.

The CCA plots for both depths showed the Arctic stations driven away by the *L. minimus* distribution but also many Mediterranean and North Atlantic stations related to *L. hargravesii.*

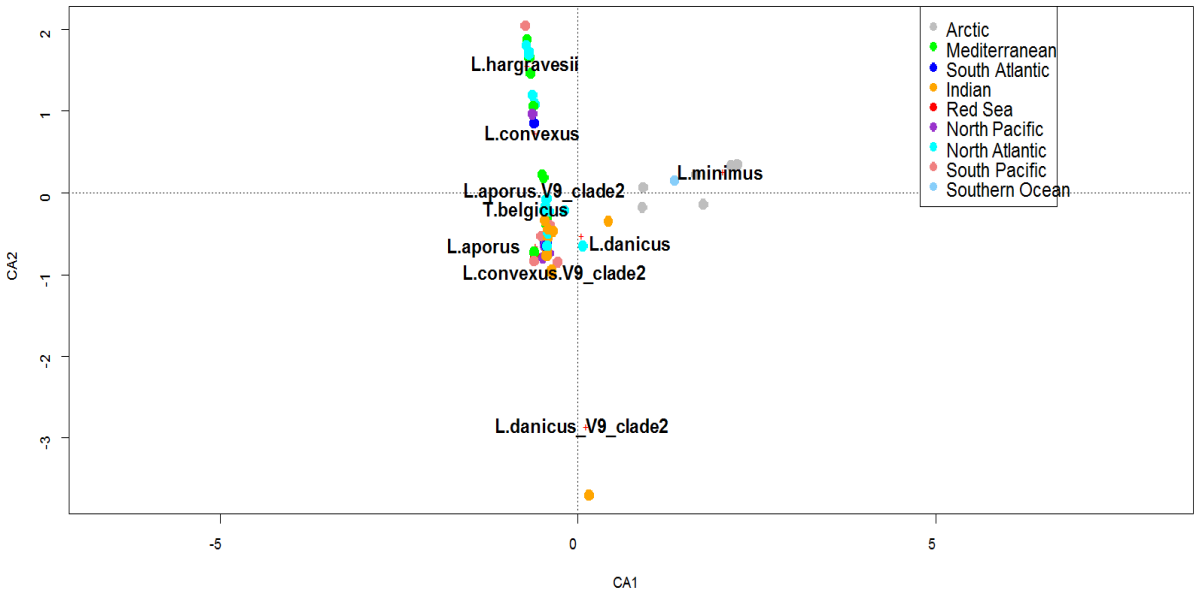


Figure 5.3.3.9 CCA plot of Leptocylindraceae rarefied abundances from the Tara surface 5-20 µm size fraction. The stations are colour-labelled based on their geographical position.

In surface Tara samples the *L. danicus* V9 clade 2 placed away an Indian and a South Pacific station (overlapped by the Indian station on the graph) while in DCM samples *L. aporus* did the same for a South Pacific station.

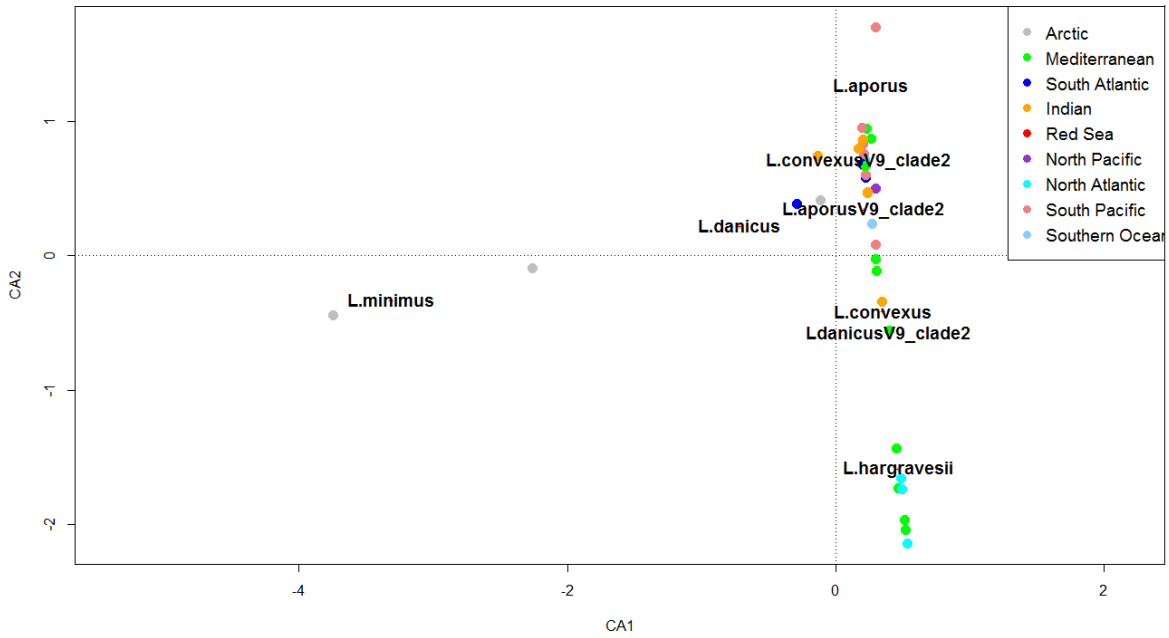
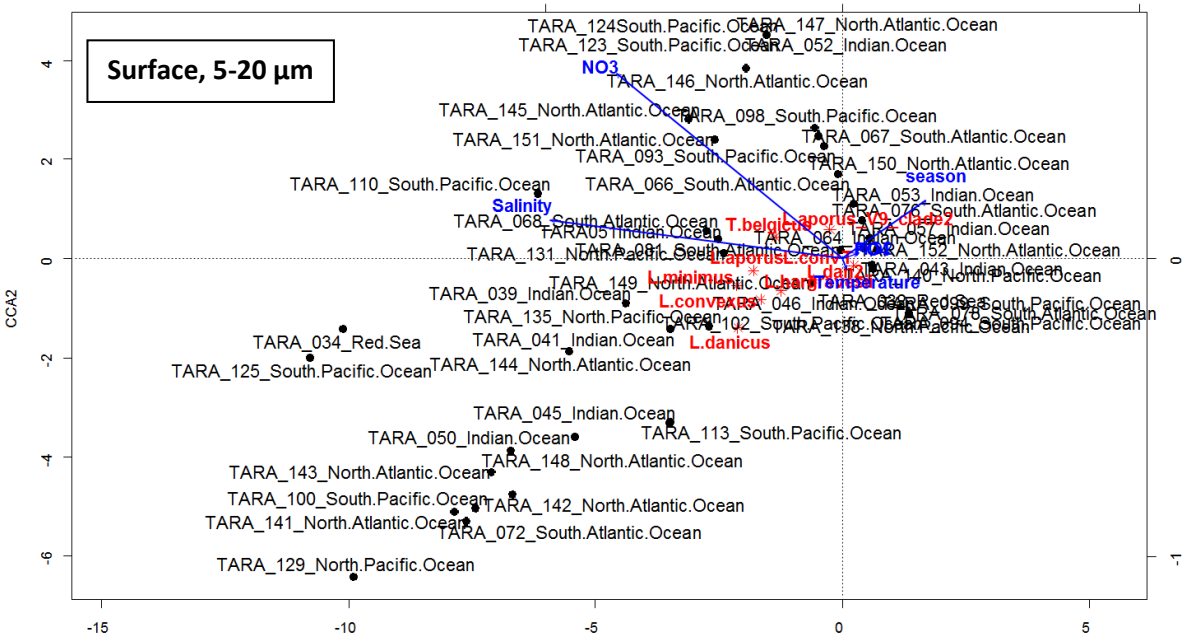


Figure 5.3.3.10 CCA plot of the Tara DCM 5-20 µm size fraction Leptocylindraceae rarified abundances. The stations are colour labelled based on geographical position.

Environmental data

Through the bioenv function the environmental variables that best correlated with the similarities of the biological community in the samples (sequence abundance) were temperature, salinity, NO2, NO3, PO4 and season for surface with a Spearman correlation of 0.153 and temperature, salinity, NO2, PO4, SiO2 and season for DCM with a Spearman correlation of 0.092. The CCA analysis on the surface samples using variables selected with bioenv showed no clear geographical pattern; the axes explained 22.3% of the total variance (Fig. 5.3.3.11). On the other hand, 63.9% of the variance in the DCM samples was explained; indeed, a distinct group of Mediterranean and North Atlantic stations and a group of Indian, South Pacific and South Atlantic stations were easily observed.



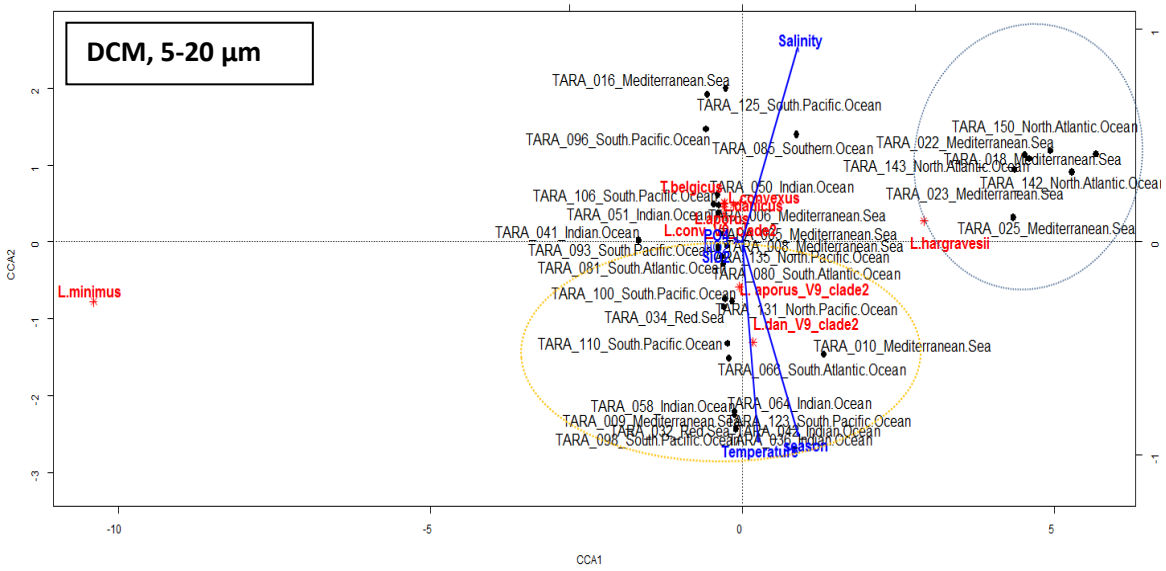
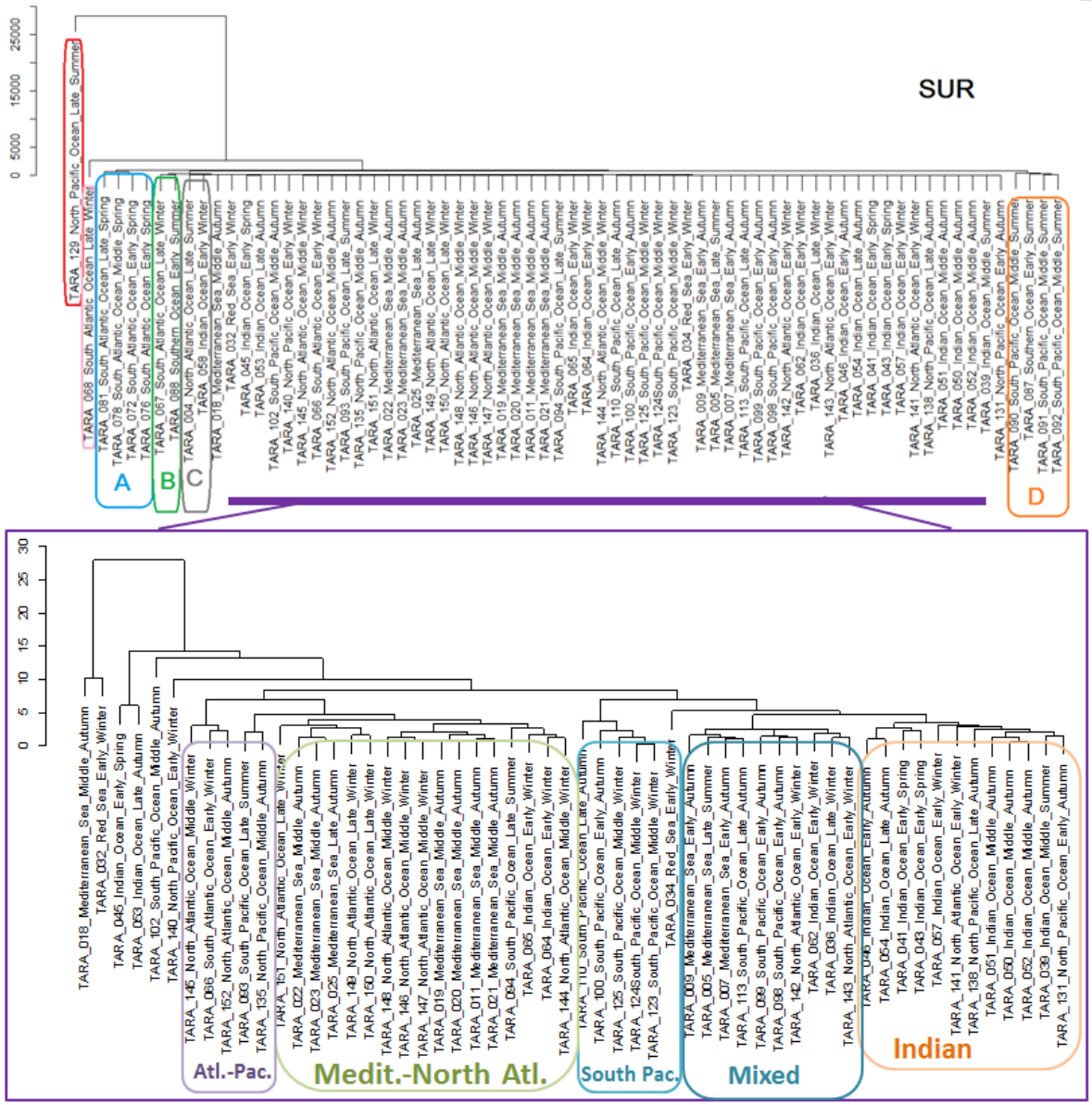


Figure 5.3.3.11 CCA analysis of surface (above) and DCM (below) Tara Leptocylindraceae samples, 5-20 µm size fraction, and selected environmental parameters. The Mediterranean/ North Atlantic (blue circle) and the Indian/South Pacific/South Atlantic group (orange circle) are highlighted. The Arctic stations, two stations in the Mediterranean Sea and one North Atlantic station are not included. In surface samples, the axes explained 22.3% of total variance whereas in DCM 63.9%.

Environmental data were missing for all Arctic, two Mediterranean and one North Atlantic stations. The available environmental parameters clustered mainly based on geographical position in the HCA analysis (Fig. 5.3.3.12). A few stations were very different from the majority and these were mainly South Atlantic stations for both depths plus an Indian and some South Pacific stations for the surface samples. The rest of the stations clustered in five main clades, the Atlantic-Pacific one, the Mediterranean/ North Atlantic, the South Pacific, the mixed clade and the Indian one. The same clusters were maintained in both depths with only the Indian and South Pacific clades splitted in the DCM samples.


SUR
A
B
C
D
Atl.-Pac.
Medit.-North Atl.
South Pac.
Mixed
Indian

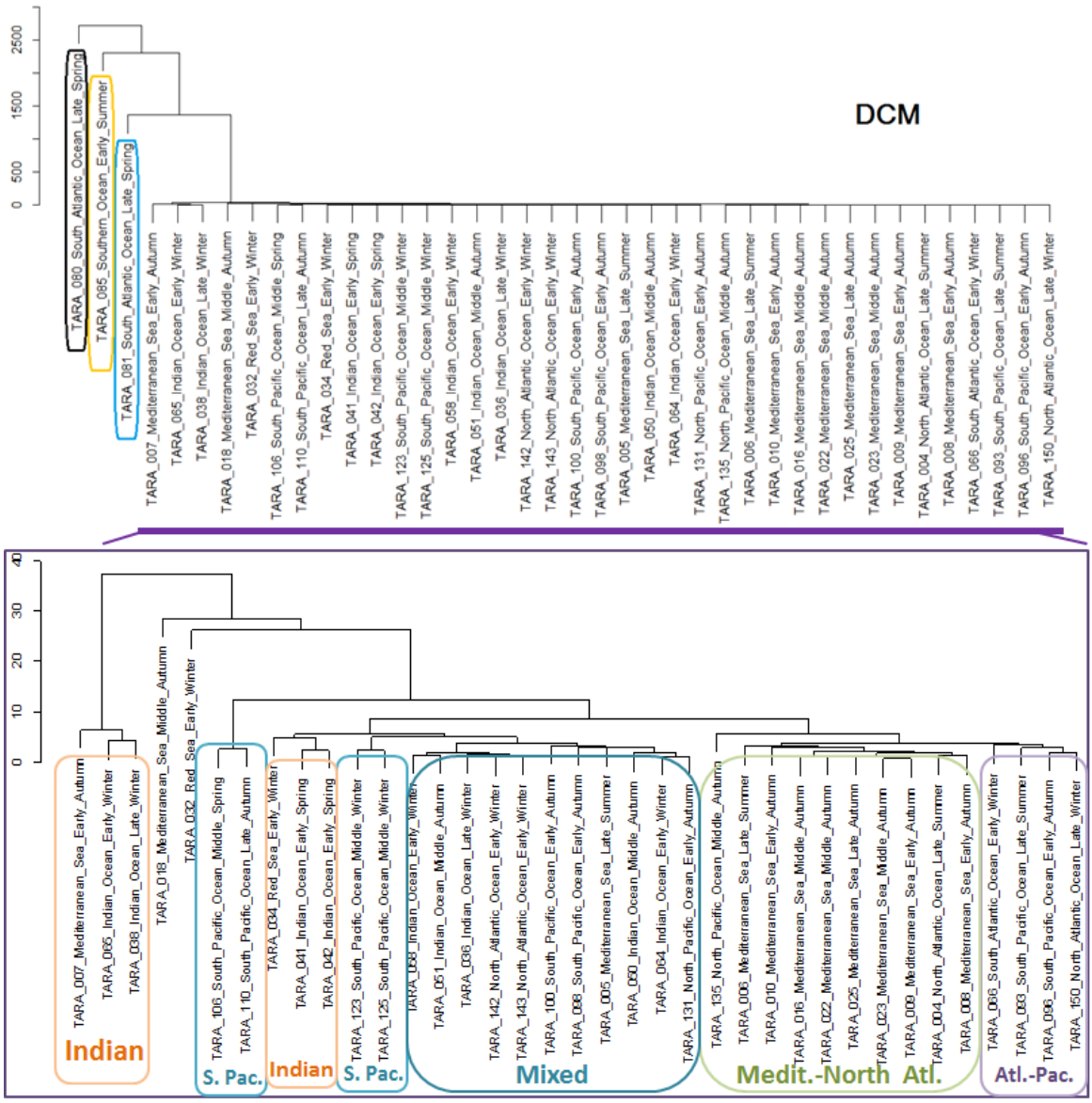


Figure 5.3.3.12 HCA plots of the Tara surface and DCM, 5-20 µm size fraction environmental parameters, without the Arctic stations, two stations in the Mediterranean Sea and one North Atlantic station.

The CCA analysis on the same dataset confirmed the South Atlantic stations to greatly differ from the rest due to SiO2 in surface and SiO2 and NO2 in DCM (Fig. 5.3.3.13 and 14). In surface, certain South Pacific stations were different based on PO4 and NO2. In DCM the TARA_085 Southern Ocean station was distinct due to PO4 and the TARA_032 Red Sea and TARA_018 Mediterranean due to NO3.

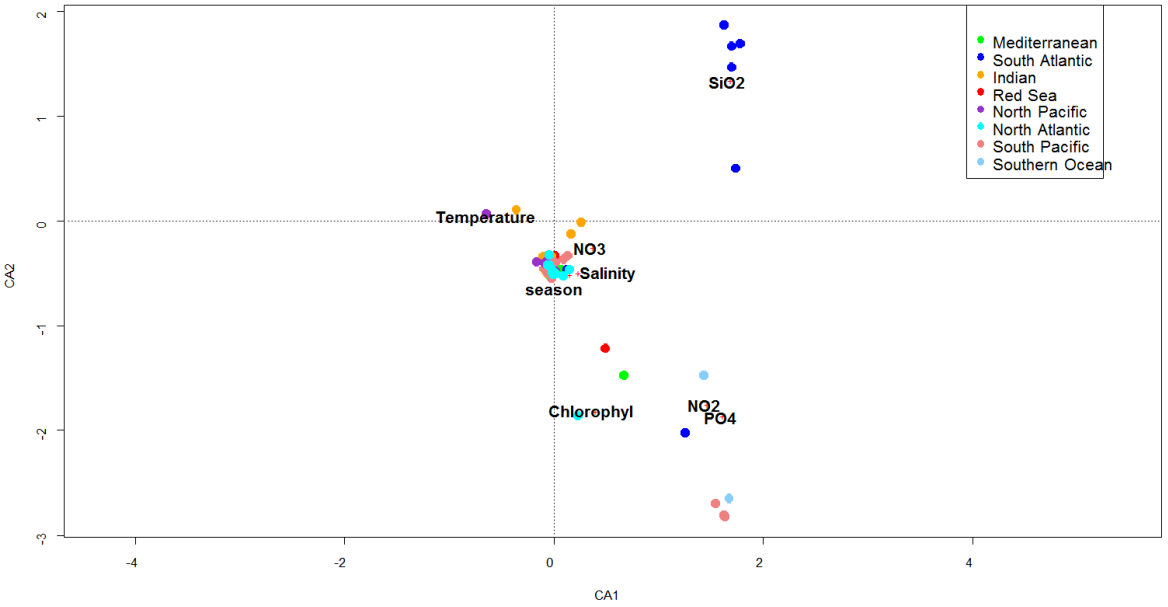


Figure 5.3.3.13 CCA plot of the Tara surface, 5-20 µm size fraction environmental parameters, without the Arctic stations, two stations in the Mediterranean Sea and one North Atlantic station.

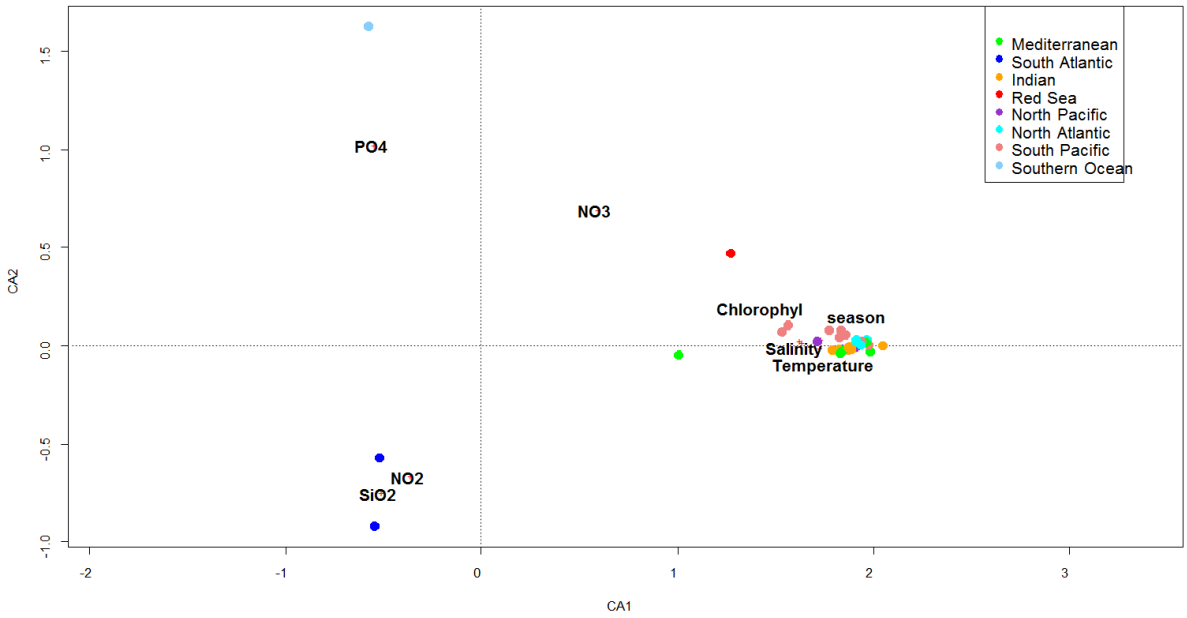


Figure 5.3.3.14 CCA plot of the Tara DCM, 5-20 µm size fraction environmental parameters, without the Arctic stations.

5.4. Discussion

5.4.1. Leptocylindraceae diversity

In the phylogenetic tree of the Leptocylindraceae family established by Nanjappa et al. (2013) two main clades were recognized in *Leptocylindrus,* one with *L. convexus* closer to *L. minimus* and then to *L. aporus*, and the other with *L. danicus* grouping with *L. hargravesii*, while *T. belgicus* was far from all being a completely different genus. All the phylogenetic trees produced in the current

study, based either on the V4 and V9 LTER-MC dataset or the worldwide V9 dataset, showed the same topology with well supported species' clades. What was different and detected for the first time here was certain clades within *L. danicus, L. aporus, L. hargravesii* and *L. convexus,* a result that showed an unexpected intraspecific diversity of the family.

To start with, the use of a recent technology such as HTS metabarcoding, in addition to light microscopy, proved to be a valuable complementary method to assess species diversity and abundance at the LTER-MC site. In fact, much more information was added on the seasonality of all Leptocylindraceae species while at the same time it became apparent that classical approaches, such as counting in light microscopy, can help in the interpretation of HTS metabarcoding results. In particular, diversity recovered at individual MC samples with HTS-metabarcoding was higher than with microscopy because of the difficulty in identification of certain *Leptocylindrus* species under the microscope and due to the larger volumes of water filtered and examined in the HTS sequencing procedure, which allowed for a better coverage of the rare species on each date. Nevertheless, the relative abundance of light microscopy counts and HTS reads over the whole diatom datasets were comparable, except for a few dates when HTS reads were higher, probably because of the low representation of Leptocylindraceae in the total diatoms, like on 03/04/2012, or dates when microscopy counts were higher due to the opposite situation, like on 28/10/2013. The match between the two datasets was not a big surprise since it has been already shown in several studies that HTS counts can provide a reliable estimate of diatom relative abundance, especially at the genus level, in a given sample (Piredda et al., 2016; Malviya et al., 2016). However, in HTS metabarcoding analysis, diversity estimates could be also affected by any possible variation among individuals (Behnke et al., 2004). This bias though can be excluded in this case, since ribotypes of the named clades can be met in more than one year or stations, implying intraspecific variation rather than intraindividual. Studies so far have not shown rDNA copy number variation to be a major concern for diatoms, in contrast to e.g. dinoflagellates, where the presence of high copy numbers is particularly pronounced (Godhe et al., 2008; Malviya et al., 2016; Piredda et al., 2016). Furthermore, extremely rare species or taxa at certain regions

and/or seasons could be the reason for their apparent absence in HTS dataset. In the MC dataset, the high correspondence between HTS and microscopy for Leptocylindraceae implies that the mentioned biases are not a problem for the diversity and distribution assessment of this diatom family.

Another HTS bias to be discussed here is the sequencing artifacts such as nucleotide calling errors, which can artificially inflate biodiversity estimates. Therefore, artifact removal processes are usually employed, as it was done in the current analysis by quality filtering and preprocessing and removal of singletons (MC and Tara dataset) and doubletons (MC dataset). The problem that comes up at this step is mostly related to the rare taxa, which are represented by low-abundance sequences sensitive to artifact removal. Several studies have shown that low abundance sequences such as singletons, doubletons and tripletons, which are usually discarded during sequence filtering processes due to their lower quality, may be informative and valuable in reflecting rare and/or unique lineages in communities (Kauserud et al., 2012; Zhan et al., 2013; Zhan et al., 2014). Zhan et al. (2014) demonstrated that elimination of rare taxa occurred at all filtering stringencies examined with more rare taxa eliminated as stringency increased. The same authors suggest the use of internal (reliable operational taxonomic units selected from communities themselves) and external (indicator species spiked into communities) references as a practical strategy to evaluate artifact removal process. The limited knowledge on biodiversity in complex communities, and thus lack of suitable references for eliminating errors/ artifacts while preserving real sequences (Kunin et al., 2010; Bowen et al., 2012; Edgar, 2013), makes the quality filtering process even harder and an immense technical challenge that should be resolved.

On the other hand, in many cases rarity has been chosen to be ignored based on some evidences that the rare biosphere is largely composed of a long queue of errors (Behnke et al., 2011; Junemann et al., 2013; Kunin et al., 2010). The latter approach has been chosen in the current study, where the risk of losing “informative” reads has been preferred to that of introducing potential errors; the thresholds used during the quality filtering have been considered to be sufficient for detecting diversity and distribution at a satisfying and safe degree for the purposes

of this study, which is not the exploration of the rare taxa. In fact, none of the two sequencings (MC and Tara) was aiming at rare taxa, which may have been missed when they were close to the detection limit of the specific HTS sequencing depth, or in samples where other species were markedly dominant. In the future, the removed sequences can be explored by careful by-eye observation of their alignment with Sanger sequences, while a different experimental design sould have been followed in a study focusing on the rarity in the community.

As already stated in the Introduction, preferential PCR amplification (Wintzingerode et al., 1997) or high copy numbers of V4 and/ or V9 could also distort the actual proportion among the species in a sample. As stated before, rDNA copy number variation has not been found to be a major problem for diatoms (Godhe et al., 2008; Malviya et al., 2016; Piredda et al., 2016). In the present study on Leptocylindraceae, there were very few inconsistencies between the two methodologies used, HTS metabarcoding and light microscopy, mainly when numbers of sequences were lower, probably because of higher errors of both methods in assessing abundances when record numbers are low. In addition, high number of sequences and ribotypes were limited to specific dates, suggesting blooms of selected taxa rather than high copy numbers effect. Using deep sequencing and utilizing more than one marker is a useful way to bypass biases resulting from DNA metabarcoding, such as detection limit for specimens with low biomass and primer specificity (Elbrecht and Leese, 2015). When it comes to Tara stations, sampling in different seasons could cover a wider range of species including those which could have been too rare to be detected during one single sampling date.

The two markers used in this study worked equally well since both of them were able to detect all species with a perfect overlap for many cases while at the same time they seemed to compensate each other, as they both revealed unique intraspecific clades. The uniformity of the V9 swarm OTUs compared to the V4 ones was an obvious sign that V9 was able to detect much less diversity than V4 but still had the power to differentiate populations on a level that V4 failed to do so. In more details, V4 worked better for *L. danicus* detecting *L. danicus* V4 clade 2 for the first time in LTER-MC dataset in this study and separating *L. hargravesii* in a more straightforward way than V9

did. In addition it was able to detect the DEL clade within *L. hargravesii.* On the other hand, *L danicus* in LTER-MC dataset was less variable based on V9 but a separate clade (V9 clade 2) was detected in Tara and BioMarKs dataset which was absent in LTER-MC dataset. *L. hargravesii* was not clearly separated from *L. danicus* with V9 (only one nucleotide difference) neither any intraspecific variability was detected in the LTER-MC nor the worldwide dataset, but higher sequence numbers were noted with V9 than with V4 in LTER-MC. This was also the case for *T. belgicus.* An intraspecific clade was also detected for *L. convexus* in the V9 worldwide dataset which was absent in LTER-MC. Finally, *L. aporus* was found more clearly diverse by V9 than by V4 with the identification of *L. aporus* V9 clade 2 in both datasets. These discrepancies between markers could be explained by the V4 and V9 variable nature. Indeed V4 is the longest variable region in the rRNA gene (350 to 450 nt) and has the highest length polymorphism (Hadziavdic et al., 2014). V9 was also characterized by high nucleotide variability in the center of the region, much shorter though compared to V4 (approximately 60 nt) which makes V4 a better marker to access diversity (Dunthorn et al., 2012; Hadziavdic et al., 2014). The *L. hargravesii* grouping within *L. danicus* but also the *L. hargravesii* V4 DEL clade that remains unnoticed in V9 confirms the lower resolution power of V9. However, V9 should not be discarded as a choice for HTS DNA metabarcoding since there might be cases that good results can be obtained like it happened here. The suitability of each marker for detecting diversity depends on the species to a great degree as each species follows each own evolutionary story. At the same time it has been proven that primer regions within the 18s rRNA gene are quite different regarding their universality (Hadziavdic et al., 2014) which could make V4 or V9 better choice for different genera. In fact, the V4 and V9 primers were checked regarding their match on each Leptocylindraceae species and it was noted that all species have a higher affinity for the V9 primers. This difference could explain why V9 might be a better choice for detecting a higher number of sequences in *L. hargravesii, L. convexus* and *T. belgicus*.

The variation found in V4 and V9 markers might correspond to higher differences in more variable markers such as ITS, which could mean that the clades identified within species might be more

diverse than revealed here. Actually, it cannot be excluded that the different clades found within the species correspond even to separate species, considering that known species (*L. danicus* and *L. hargravesii*) were also found inside the same OTU or phylogenetic clade due to their high similarity in V9. In the worldwide dataset, *L. minimus* was found to be a quite variable group. Four out of the ten OTUs in the swarm analysis were composed by *L. minimus* sequences but it should be noted that the number of sequences of these *L. minimus'* OTUs (except the main one, OTU#4) were much lower than the rest of the species and particularly these clades had sequences even as low as 34 compared to other species' clades e.g. *L. danicus* V9 clade 2, which consisted of 11,352 sequences. So even though the *L. danicus* clade 2 was based only on one nucleotide difference in a conserved region within *L. danicus*, its actual existence is much more supported by the high number of occurrences than the *L. minimus* small clades. Therefore, these clades in *L. minimus* should be treated with caution. Finally, the Baffin Bay clade, possibly a yet undescribed taxon, detected in the V4 BioMarKs dataset by Nanjappa et al. (2014a) does not seem to have a V9 signature in terms of sequence diversity compared to the reference sequences. The *L. danicus* V9 clade 2 was further explored in the BioMarKs dataset and the majority of the sequences were located in the Oslo Fjord, where the Baffin Bay clade was also abundant in Nanjappa et al. (2014a). However, *L. danicus, L. hargravesii, L. minimus* and *T. belgicus* were also more abundant in the Oslo Fjord compared to the rest of the BioMarKs stations, so the *L. danicus* V9 clade 2 spatial pattern could be a result of a bloom event in this period rather than an expression of the locality of the Baffin Bay clade. Further investigation, with isolation of strains of the species from different areas, is needed in order to clarify all the doubtful cases described above.

5.4.2. Leptocylindraceae distribution in time and space

Based on the so far available knowledge and information on Leptocylindraceae family in GoN, *L. danicus* was considered to be present from late autumn through mid-summer (retrieved from mid-November to mid-July), *L. aporus* was identified as the species blooming in summer but also present in autumn (retrieved from mid-July to mid-November), *L. hargravesii* and *L.* convexus were considered to be rare and have the narrowest temporal distribution, only in winter months

(*L. hargravesii* retrieved in December and January; *L. convexus* retrieved from January to end of March and observed in April samples). Finally *T. belgicus* was found from late summer through the autumn (retrieved from end of August to beginning of November) (Nanjappa et al., 2013). Comparing the HTS numbers with the corresponding seasonality resulting from the LTER-MC microscopic counts, obvious differences could be noticed in the current analysis. Microscopy detected a narrower temporal window than the actual one represented by HTS, which was expected considering the higher sensitivity of the latter method. For rare species and species hard to discriminate under the microscope, HTS was a very important complementary method in order to get the real picture of their abundance. At a larger geographic scale, i.e. based on BioMarKs data and sequences found in GenBank, *L. danicus* was considered to be the more abundant species with *L. aporus* following and then *L. convexus,* while *L. minimus* and *T. belgicus* were of similar abundance (Nanjappa et al., 2014a). Still in Nanjappa et al. (2014a), *L. aporus* has also been considered the most widespread species; *L. convexus, L. danicus* and *L. hargravesii* were also widespread, with the second one having been detected in more stations than the latter; *T. belgicus* and *L. minimus* were absent from most stations with the latter being restricted to colder European waters and in the Black Sea. The present results mainly confirmed both the temporal and spatial distribution, especially for *L. danicus* and *L. aporus,* but also added important information on the distribution of the less studied *L. convexus, L. hargravesii, L. minimus* and *T. belgicus* as well as of the new populations identified within each species. The seasonal patterns were expanded for all species and clarified for the new populations while their specific geographical patterns could be identified supporting the partition of phytoplankton in the seas regarding temperature adaptation into the temperate and the tropical class.

To discuss temporal distribution in more detail, the DNA metabarcoding analysis results of both V4 and V9 on Leptocylindraceae at the LTER-MC station showed indeed a consistently high abundance of *L. aporus* during summer and beginning of autumn (July-September) and a bimodal increase of *L. danicus* during spring and autumn (April-October) through all three years. In addition, the new data led to a more complete picture of the species seasonality, the

understanding of which so far was based on very limited information. So what was new here, compared to Nanjappa's (2012) assessment of the species seasonality, was the presence of *L. aporus* also in June and May and *L. danicus* in September. But even more interesting was the different seasonality of the *L. aporus* V9 clade 2 compared to the main *L. aporus* clade. The V9 clade 2 showed no summer peak in any of the three years but rather a big one in October 2013 and two smaller ones in February-March of all years and in September 2013. Although the environmental parameters between July/ August and September/ October generally do not differ in a great degree, February and May or October do, a fact that makes the small *L. aporus* V9 clade 2 population, if not an adapted "non-summer" population, a more plastic one regarding its tolerance to colder conditions. On the other hand, the *L. danicus* V4 clade 2 was present in all three years following a similar pattern as the main *L. danicus* clade with the addition of only few lower peaks. The most important would be the one of December 2013 and then February 2012 and March 2011. Although the evidence is not as strong as in the *L. aporus* case, the *L. danicus* V4 clade 2 population could also be characterized by higher plasticity than the rest of *L. danicus.*

Coming back to *L. aporus,* in BioMarKs and Tara dataset this time, the species was indeed the most abundant and highly widespread, found in stations of all longitudes and latitudes. A striking result that adds up to what has just been described based on the LTER-MC dataset was the distribution and the abundance of the two *L. aporus* clades. The *L. aporus* V9 clade 1 was the most abundant at some stations while the V9 clade 2 was the most widespread one in the surface samples. Yet, in the deep chlorophyll maximum samples the situation changed with *L. aporus* V9 clade 2 being both the most abundant in all stations and widespread. At the same time, the stations where clade 1 was detected were in the Mediterranean, the Red Sea and the South Pacific Ocean, all of them close to the equator. Stations with the highest abundance for this clade were in Mediterranean Sea. *Leptocylindrus aporus* V9 clade 2 was found at the same stations as the main clade but also in the Arabian Sea, and at more stations in the Indian Ocean close to South Africa, Southern Atlantic and South Pacific Ocean. The station with the highest abundance was again in Mediterranean but other stations with high abundance were located close to South Africa and

South Pacific. Finally, *L. aporus* clade 2 was also present in Arctic stations where the main clade was completely absent. These facts support the hypothesis stated already about clade 2 being a more plastic or diverse population that can grow equally well in temperate, tropical and polar zones but also in deep ocean conditions. In LTER-MC, this more plastic clade 2 could have slightly readjusted its seasonal pattern through selection and have now occupied a different temporal niche compared to the main *L. aporus* clade, thriving in the new niche since it seems that *L. aporus* clade 1 is a better competitor when they both co-exist in a favorable environment. The *L. aporus* V9 clade 2 could have become adapted to these conditions and therefore now be a completely different population. This hypothesis could be answered with certainty only by population genetics.

Continuing with the worldwide distribution, the *L. danicus* V9 clade 2 did not diversify greatly from the main *L. danicus* regarding its spatial pattern. *L. danicus* V9 clade 2 was present in the same stations that the main *L. danicus* clade was too, but with a higher preference for the subtropic and tropic zone stations. When looking into the CCA analysis of surface samples but also DCM, the *L. danicus* V9 clade 2 population was plotted further than the main *L. danicus* population. The stations where *L. danicus* V9 clade 2 was detected were in the Indian and South Pacific Ocean. The *L. danicus* V9 clade 2 population, which is much smaller than the *L. aporus* V9 clade 2 population, might be currently diversifying and moving slowly to more tropical environments following the same strategy as the *L. aporus* V9 clade 2. The fact that is a very recent clade would also explain why it has not been detected in the LTER-MC station yet.

The next revelation of the HTS analysis was *L. hargravesii* both seasonal and spatial distribution. The species *L. hargravesii* had previously been detected only in January but in this study it was actually much more abundant during summer – beginning of autumn (July - September) escaping from its characterization as a strictly winter species by the evidences so far only based on observations on isolated strains (Nanjappa, 2012). This is not a huge surprise though since *L. hargravesii* is a rare species and therefore hard to “catch” during isolations. After this indication based on HTS, isolations during end of August- beginning of September 2014 targeting *L.*

hargravesii were performed and the molecular identification did indeed lead to the confirmation of the metabarcoding analysis results. The possibility of the *L. hargravesii* DEL clade to be a rare intraindividual variant within individuals also possessing the normal V4 type has been excluded, because the deletion was not the only difference between the two clades; there were 12 more mismatches in conserved domains within Leptocylindraceae. In addition, the ribotypes forming the DEL clade were found in all three years, indicating that it is a small population of a certain variant rather than a case of intraindividual variation. The *L. hargravesii* DEL clade showed a preference for July so it could be a population better adapted to this period. In the worldwide dataset, *L. hargravesii* showed a more extended distribution in the Mediterranean Sea than *L. danicus* but a more limited one globally since *L. danicus* was also detected in temperate and Arctic stations. Most *L. hargravesii* sequences from surface samples were detected in temperate/subtropical regions such as the Mediterranean Sea but DCM sequences were also present in the Indian Ocean station near Port Elizabeth, south of Africa. In fact, while *L. hargravesii* was less abundant than *L. danicus* in the surface samples, it reached higher numbers in the DCM samples compared to *L. danicus.* The expansion of the seasonal distribution of *L. hargravesii* in summer showed that the species is more flexible than thought while the particular geographical distribution showed a subtropic specification but also a deep sea preference which could explain the rarity of the species during strain isolations in LTER-MC.

Important information was added for *L. convexus* distribution as well. The species was found quite abundant and took part mainly in the July bloom peak in all three years. So although being present also in winter, the species escaped the strict characterization as a winter species by the evidence so far available based on strain isolation (Nanjappa, 2012) and proved to be rather a spring and summer species. In the worldwide dataset, *L. convexus* was the second most abundant Leptocylindraceae species and one of the most widespread as well, while *L. danicus* was found to be the third more abundant species. But in contrast to the other two worldwide-distributed species, *L. aporus* and *L. danicus, L. convexus* was not found in any Arctic stations. *L. convexus* main clade was mainly present in the Mediterranean Sea while stations with intermediate

abundance were located in South Africa, South Atlantic and South Pacific Ocean. The *L. convexus* V9 clade 2 was also present in Mediterranean sites but it was more widespread worldwide, found also in more stations in the Red Sea, the Arabian Sea and in additional sites close to South Africa and South America in high abundances. The CCA plots of DCM and surface samples show a slight divergence between the two populations driven by the Mediterranean stations for the main clade and the South Pacific and Indian Ocean for the V9 clade 2 population. The abundances were conserved in the surface and DCM samples, with *L. convexus* V9 clade 2 dominated at both depths. Additionally, *L. convexus* V9 clade 2 was the most widespread *L. convexus* population. Therefore, in *L. convexus* there is a more plastic population just like in *L. aporus,* able to adapt to temperate and tropical zones and deep sea. The *L. convexus* V9 clade 2 population was absent from the LTER-MC station. Here it should also be noted that the main populations and the clades found within *L. aporus* and *L. convexus* show a very high resemblance regarding their world distribution which is justifiable considering the two species are genetically closely related and both summer occurring species so the chances for their physiological responses to resemble are higher. Niche conservation is also presupposed here.

L. minimus is a species that has never been found in GoN and the results of the current study confirmed that. One of the most striking observations in the Tara world maps was the bipolar distribution of *L. minimus*, which dominated the arctic and Antarctic stations with two single exceptions of a North Atlantic and an Indian Ocean station. The North Atlantic station was sampled in winter and the temperature was around 14°C while the Indian one was sampled in early spring when the temperature was as high as 29°C. Although the Indian population was constituted only by five ribotypes, four of them represented by only a single sequence, it is a result that cannot be ignored and implies a possible plasticity of *L. minimus* that allows it to survive in high temperatures under specific circumstances which are not clear yet. Nevertheless, the higher abundance in the rest of the stations showed that this species prefers arctic and/or antarctic environments which means e.g. very low temperatures and that is why it is so rare in the temperate and tropical regions, including GoN where it is completely absent. This bipolar pattern

is strongly supported by the HCA and CCA results. In all the analyses the arctic stations were placed further than the rest of the stations and they were always linked with *L. minimus.* This finding is essential to understand the ecology of this species which was quite recently identified as a separate species different from *T. belgicus* (the last one shows a completely different geographical pattern - discussed right after - further supporting their separation).

Tenuicylindrus belgicus showed an expanded seasonal distribution with a significant high number of sequences also in December, beyond the August - October period. It was the only species with almost zero presence in spring, a result that could be also influenced by the high presence of the rest of the species in the environmental sample sent for sequencing. *Tenuicylindrus belgicus* was also the sparsest species worldwide, restricted mainly to the Mediterranean; the Indian station showed a very low abundance of just three individuals in the surface samples. *Tenuicylindrus belgicus* was the only species of the family that was so localized which again could be an artifact of the sequencing process. Yet again, the low numbers or complete absence in most seasons and stations reveal a species with a quite limited capability in terms of adaptation to many different environments. Therefore, despite the fact that *Tenuicylindrus* morphologically looks close to the rest of the Leptocylindraceae family, not only it is a genetically different genus but it also shows a very different behavior in terms of seasonal and geographical distribution.

Summing up on the family spatial distribution, it seems that all *Leptocylindrus* species, except *L. minimus,* are divided in two main geographic occurrences: one consisted of mostly tempered and subtropic zone sites and the other was broader including also tropical and arctic sites. This finding is consistent with Hulburt's (1982) conclusion that two categories of phytoplankton occur in the sea regarding temperature adaptation; the temperate class of the species that grow at temperatures ranging from approximately 2°C to 25°C and the tropical class which has a growth range from around 12 to 34°C. The season of sampling could be a factor for Leptocylindraceae, probably more notably for *L. hargravesii*, but it cannot be considered the sole responsible for the spatial patterns recorded in this study since species found in stations sampled in certain seasons were also absent in others sampled in the same season. The geographical distribution discussed

here shows clearly a connection to species and even to clades.

Looking at the Leptocylindraceae family as a whole, it showed a structure driven by each species preference for certain environments rather than clearly seasonal clustering. That was why HCA and CCA analysis at LTER-MC station did not show a clear clustering of months based on seasons but based on the months each species blooms. So, for example July and September, the months when *L. aporus* blooms, were placed together while the same happened for April and October, when *L. danicus* blooms. It could be assumed that V9 is a better marker for the exploration of the family structure since V9 detects higher numbers of sequences of the species increasing the corresponding seasonal coverage. The specific clades e.g. *L. hargravesii* DEL clade which were only detected in V4 imply that, as for diversity and distribution, also for community structure both markers should be used.

At the worldwide scale, the Leptocylindraceae family structure in both depths showed a clustering based on geographic proximity although there were certain stations distinguishing from the rest or placed in mixed clusters. The majority of Leptocylindraceae was detected in the 5 – 20 µm size fraction, in coastal and surface samples as expected considering the size and the ecology of the species. The geographical distribution analysis was performed on two different groups though; the DCM 5-20 µm and the surface 5 – 20 µm. This was done to be sure that the sampling procedure was, if not exactly the same, at least quite similar for all stations. That is why BioMarKs stations were also removed. In the end, we must be sure that the geographical pattern that we saw was a result of the actual distribution of the species rather than a bias, product of the different methodologies used in each project.

It should be reminded here again that Tara stations were sampled in different seasons, a fact that might have influenced the actual composition of the diatom community. A possible effect of this could have been the *L. hargravesii* high abundance in Mediterranean, which was sampled in summer and autumn. The CCA analysis showed a grouping of some stations related to *L. hargravesii* which were actually sampled in winter and autumn. When looking into the specific corresponding stations they were Mediterranean, North Atlantic and South Pacific stations where

L. hargravesii was detected. Yet there were still stations sampled in the autumn and winter, such as in the Indian Ocean, where *L. danicus* was dominant and *L. hargravesii* was almost absent. So even though the time of sampling could have been a factor for the prevalence of *L. hargravesii* over *L. danicus* in certain stations, the geographical position and therefore the related areal characteristics of the temperate stations had no less importance.

The environmental parameters of the available Tara stations showed that in both surface and DCM, PO4 and NO2 drive certain South Pacific, South Atlantic and Southern Ocean stations away from the rest and SiO2 drives certain South Atlantic stations away. Temperature and season were not major drivers of any specific structure. The South Atlantic group includes stations such as 068 and 078 which are mentioned to be part of the Agulhas rings (http://www.igs.cnrs-mrs.fr/Tara_Agulhas/). Agulhas rings are a result of the Agulhas Current coming from southwest Indian Ocean and reaching the South African east coast which causes large scale cyclonic meanders known as Natal pulses (Leeuwen et al., 2000). The rings are created at the tip of South Africa (Station_068) and slowly drift across the entire South Atlantic towards the coast of Brazil (Station_078). It has been already shown within the Tara expedition research that the Agulhas rings drive complex nitrogen cycling and can shift the plankton diversity between the Indian and Atlantic Ocean (Villar et al., 2015). So it is indeed a region that diversifies from the rest and might have influenced also the species of interest of this study.

The limited seasonal coverage of the Tara sampling is the more fundamental problem in this dataset, but as far as the relationship between spatial/temporal distribution of individual species and environmental variables is concerned, it is hard to draw any solid conclusion since Leptocylindraceae species are a part, and often a minor part, of the whole phytoplankton communities. Indeed the effort of analyzing these relationships did not lead to a clear picture. Temperature at LTER-MC and temperature, salinity, selected nutrients and seasons at Tara stations were the factors that best correlated with the sample similarities of the biological community (in terms of species abundance). So temperature and season were factors that seem to matter to a great extent. The low values of the correlation coefficients could be explained by

the fact that the Leptocylindraceae community is under the influence of other environmental parameters, not included in the measured ones, as well as by interactions with other diatom and other plankton species. In addition, when abundance and environmental parameters were tested separately there were some obvious violations of specific patterns, implying a noise in the dataset that could be overcome with the extension of the analysis to multiple years on similar dates. However, the DCM community appeared to be more influenced by the factors examined (x, y), possibly because the competition at higher depths is less while the environment is more stable. Nonetheless, the outcome of the environmental analysis on both LTER-MC and Tara stations confirmed that temperature was an important factor among the studied ones, for the temporal and spatial distribution of the Leptocylindraceae family.

5.4.3. Conclusion

A point by point comparison with the recent results of a similar study on Leptocylindraceae across European waters (Nanjappa et al., 2014a; specific points mentioned in the end of section 5.1.2) can be done here:

1. The species diversity in the Leptocylindraceae is indeed low but higher than known before this study.
2. *L. danicus* and *L. hargravesii* were resolved in V4, both phylogenetic tree and Swarm, while in V9 *L. hargravesii* was incorporated as a *L. danicus* population. Despite that it is possible to identify each species based on a single nucleotide base difference.
3. *L. aporus* dominated the V4 but also the V9 dataset.
4. V4 produced more reliable distance trees compared to the lower resolution offered by the shorter V9 region. Still, there can be specific intraspecific diversity that only V9 can detect such as clade 2 in *L. aporus* clade.
5. V9 showed a higher detection power of Leptocylindraceae, especially in the case of *T. belgicus* and *L. hargravesii*. In contrast to Nanjappa et al., (2014a) the explanation of different techniques (Illumina and 454) cannot be used in the MC dataset since both markers were sequenced with Illumina. This result could be explained by the higher

variability of the V4, including the primer region, compared to V9.

6. In GoN, *L. aporus* dominated in September 2011, 2013 and July 2012 (compared to October 2009 in Nanjappa's analysis), the rest of the species were also present at those dates except *T. belgicus* in July 2012. In October 2012 and July 2013 *L. danicus* and August 2011 *L. convexus* were in higher numbers though all species were also present (compared to May 2010 in Nanjappa's analysis). Even though the months are different, the seasons are the same as Nanjappa's or slightly expanded.

In total there is agreement of results with an expansion of information thanks to the longer term nature of the last study.

Other than that, the conclusions can be summed up in the following points:

- MC dataset is much smaller than the Tara dataset and therefore a simpler one to handle. The cleaning and preprocessing analysis was also more extensive than with Tara data. Therefore the final sequences analysed for the species diversity showed much less sequencing errors. This, combined with the much smaller amount of data, made it easier to identify and define species and clades as well as their corresponding single nucleotide variations in V9. Thanks to this analysis it was eventually possible to detect similar populations in the Tara dataset. In addition, SWARM analysis was a very helpful tool for the exploration of the species diversity especially in the Tara dataset which is a vast dataset. The exploration of the phylogenetic tree would be more laborious and time consuming without any indications from SWARM OTUs and the MC dataset results.
- The intraspecific diversity is low compared to other genera but it exists. Even a single nucleotide difference can be enough to differentiate two species (e.g. *L. danicus* and *L. hargravesii* in V9) or populations (*L. aporus* V9 clade 1 and V9 clade 2). V4 and V9 complete each other regarding intraspecific diversity assessment because of the different resolution power. The different species and clades within species show specific seasonal and geographical distribution which further supports their diversity.
- The seasonality of the individual species was confirmed and most importantly expanded.

V4 and V9 complement each other in this matter due to their different detection power. The comparison with microscopy counts showed that HTS is an important complementary method for species hard to diversify morphologically or isolate due to rarity. Yet the absence of species or clades in HTS data during specific periods, especially during blooming events, should not be considered definitive since blooming ribotypes may saturate the sample and mask the presence of the rarer ones.

- The world distribution of Leptocylindraceae family revealed possible spatial patterns not only for the individual species but also for the clades identified within each species leading to the hypothesis of adapted populations.

The results from the HTS DNA metabarcoding, combined also with the physiological and transcriptomics results of the previous chapters lead to the conclusion that specific genotypes can hold an essential role in the ecology and evolution of diatoms. At this point a general question comes up, the answer to which will be important for our understanding of the entire marine environment and the marine community interactions: How specific genotypes alter the community structure and what is the impact of the phenotypic diversity offered by multiple genotypes on a given community composition (Wohlrab et al., 2016).

Chapter 6. General Conclusion and Future Perspectives

The results of the present study on the marine planktonic diatom family Leptocylindraceae have shown the many aspects of its species diversity, which otherwise show a great morphological uniformity, and thereby have answered important questions related to the seasonal and spatial distribution of the family. Eukaryotic phytoplankton has been proven to be vastly diverse in many studies so far (Falkowski and Knoll, 2007; Graham et al., 2009; Simon et al., 2009) while investigation on this field still continues (de Vargas et al., 2015). The diversity of a species is expressed in many levels, such as in the molecular level in specific markers, morphological and physiological state, niche differentiation and functional diversity. There are many different properties of a lineage (e.g. interbreeding in the biological concept, niche or adaptive zone in the ecological concept, heterogeneity in the phylogenetic concept etc.) that can be focused on in order to define a species or a population. Nevertheless, as stated by de Queiroz (2007), all properties are evidence for the characterization of a species, arising though at different times during the process of evolution. Therefore, the study of all these aspects is needed in order to fully describe how diverse a species may be and ultimately understand in what way each diversity level interacts or relates to the other. Furthermore, exploring the diversity of a species helps us also understand its spatial and temporal distribution; for example physiological and genomic adaptations associated with certain ecotypes (genetically distinct population adapted to specific environmental conditions) might maintain or create a specific distribution (Moore et al., 1998; Bibby et al., 2003; Rocap et al., 2003; Stuart et al., 2013). The present study adds important information towards this direction through the examination of **physiological, functional and molecular diversity** of Leptocylindraceae diatom species. In all three mentioned levels, a high degree of intraspecific diversity has been noticed, while indications of possibly new species were found, which was surprising since Leptocylindraceae is generally considered a species-poor family.

Starting with **physiology**, growth experiments under different temperatures of the *L. aporus* and *L. danicus*, which show a contrasting time of occurrence in the GoN, gave the first clue of intraspecific phenotypic plasticity. This physiological diversity could be due to the fact that both *L. aporus* and *L. danicus* are species with intermediate to high levels of abundance and a broad

seasonal window of occurrence; this is one of the main hypotheses of the present thesis and is met in functional and molecular diversity as well. Although strains isolated in specific season, and thus possibly adapted to the environmental temperature of this season, did not show any clear pattern indicating their temporal origin, there was a growth pattern that seemed to relate to the time strains had been kept in culture. Hence, the issue of in-culture evolution was raised, the effects of which could have led to the differentiation of at least two of the three old *L. aporus* strains and even overruled their actual reactions due to their natural state. In the recent years, there has been increasing awareness of the evolutionary potential of prolonged growth in culture (Wood et al., 2005; Lakeman et al., 2009) while evidence for genetic adaptation by phytoplankton are constantly reported (López-Rodas et al., 2008; Huertas et al., 2010; Lohbeck et al., 2012; Schlüter et al., 2014). For this reason, the in-culture evolution scenario was considered highly important and was kept in mind when assessing the results of the following analyses as well.

Moving to the **functional** diversity of *Leptocylindrus* species, inter- and intraspecific variability were both evident. First of all, the in-culture evolution effect first noticed in the physiology of *L. aporus* old strains was further supported by the significantly differentially expressed transcripts of the single old strain sequenced compared to the more recently isolated ones. Although the intrinsic character of the strain as isolated cannot be excluded as the reason for its different expression profile, the similar physiological response of another old strain during the growth experiments supports the in-culture evolution assumption. Whichever is the case, an important general conclusion of the current study is to keep in mind the age of the strains used in the experiments, which might influence the final results. If possible, an experimental plan for the appraisal of the age effect should be designed. Otherwise, newly isolated strains should be used.

The *L. aporus* expression patterns at different temperatures and the comparative analysis with the profiles of the other *Leptocylindrus* species led to the suggestion of a possible system involved in adaptation, which takes advantage of transposable elements (TE). The TE mechanism could offer the necessary intraspecific plasticity that enables individuals to cope with the constantly changing environmental conditions while in more stable environmental conditions TEs could

participate in the establishment of adapted populations. In fact, TEs are constantly gaining more attention regarding their role in stress response and evolution, establishing their involvement particularly in heat or cold shock reaction and their attraction to epigenetic changes (Ito et al., 2011; Janska et al., 2014; Lutz et al., 2015; Migicovsky and Kovalchuk, 2015; Naydenov et al., 2015; Ropars et al., 2015; Berendsen et al., 2016; Li et al., 2016; Sun et al., 2016; Zovoilis et al, 2016; Garbuz and Evgenèv, 2017). Interestingly, the stress-induced retrotransposon response has been shown to be genotype-specific in fungi and plants (Ansari et al.,2007; Long et al., 2009; Lopes et al., 2013; Berg et al., 2015). Indeed, the number of significantly differentially expressed transposable elements was different for each pair of strains in all *Leptocylindrus* species confirming a genotype specific TE profile. In addition, although TEs were a main component of the *L. aporus* response to low temperature, they were also significantly present in the differentially expressed genes between strains, while the total number of genes expressed significantly different between *L. aporus* strains were much higher than between different temperatures. This is an impressive outcome which highlights the level of intraspecific diversity in *L. aporus*. The intraspecific diversity in gene expression was also obvious in *L. danicus,* while *L. convexus* and *L. hargravesii* were more homogeneous. Comparative transcriptomics offered one more evidence in support of the hypothesis already stated in the analysis of the physiological diversity; abundant species that show a broad temporal range in GoN (*L. aporus* and *L. danicus*) were found to be more variable regarding strain gene expression compared to *L. convexus* and *L. hargravesii*, which are more scarce and have a more restricted seasonal niche. The functions that were enriched in most strains were related to membrane properties including transportation of molecules and signal transduction, which could be linked to the different capacity of each strain in responding to environmental stress or uptaking certain nutrients. Other enriched functions were related to post-translational or post-transcriptional regulation which could be again linked to the different ability of each strain to adjust its cellular metabolism and enhance resilience mechanisms under environmental challenges. The same functions that were different among strains were also found to be statistically different among species, implying that membrane properties and regulation

mechanisms have a crucial role in both intra- and interspecific diversity while TEs and the HSF regulatory system could also play a part in the plastic responses and adaptation of each species to different environmental conditions, such as temperature. It seems that the regulation of specific functions is involved in the temporal, and possibly also spatial, differentiation of the Leptocylindraceae species.

A factor that contributed greatly towards some of the valuable conclusions of this study was the use of different strains as replicates for each species in the experiments. Given the fact that multistrain studies in phytoplankton, and especially in diatoms, are almost absent while multistrain studies in most organisms have shown a significant degree of intraspecific diversity (Snyder et al., 2004; Llinás et al., 2006; Bradford et al., 2011; Dugar et al., 2013; Sobkowiak et al., 2014; Brion et al., 2015; Palma-Guerrero et al., 2016; Galinier et al., 2017), the use of different *Leptocylindrus* strains for the assessment of the species physiological and functional diversity was unconventional, yet vital for the better understanding of the ecology and evolution of the family. The results of this thesis call for further studies focused on the diversity of intraspecific responses in diatoms. Further experiments should be also designed for the clarification of the exact role of TEs, as well as for other heat/ cold stress related genes, such as SYM1. In particular, newly isolated strains should be tested under similar conditions. The acclimatization should be the same for all strains and designed as to include different time intervals, including immediate exposure to stress after the strain isolation; the gene expression for each acclimatization experiment at different time interval would provide more precise information on the actual response of each strain to perturbations.

One thing missing in this thesis, which would have helped in the interpretation of the expression results, is the genome of the strains. Even in the form of a draft genome, it would have assisted in analysing the RNA-seq data and gaining insights into the evolution of the genes of interest. Transcriptome assemblies based on a reference genome are generally considered more effective than those generated using de novo strategies alone, offering several advantages such as high sensitivity for low abundant transcripts (Martin and Wang, 2011; Marchant et al., 2016).

Nevertheless, combined, genomic and transcriptomic data help to define new genes that are not predicted with the use of an automatic annotation. In our particular case, having the genome would have been of great use in order to understand the interaction of the TEs with the surrounding region or identify any promoters involved in their activation. Additionally, a more thorough investigation of the transcripts related to the functions found interesting in the comparative transcriptomics analysis should be conducted; the focus should be on more specific pathways in order to understand the fine differences among the strains and the species. Genes related to temperature and/ or heat and cold stress could be also followed and their expression compared among the species. In any case, the gene expression profiles of each *Leptocylindrus* species need further exploration in order to fully exploit the information that they have to give us.

Lastly, interesting new information was obtained regarding the **molecular** diversity of the family, revealing unexpectedly high intraspecific variation in certain species. Genetic polymorphisms such as SNPs and INDELs were explored and this was the first time this kind of genetic markers were used in a eukaryotic phytoplankton genetic study (Rengefors et al., 2017). The genetic microvariations but also phylogenomics analysis led to a clear separation of the variable *L. danicus* and *L. aporus* from the less diverse *L. convexus* and *L. hargravesii*, which, given the similar partition highlighted by comparative gene expression analysis, comprises the third line of evidence supporting the hypothesis on species differences in the level of intraspecific diversity based on their abundance and broader or narrower seasonal niche. Despite their morphological similarity, all differences among strains and species described so far point at a considerable degree of functional and molecular variation in the Leptocylindraceae family, though this is more obvious in certain of the species than in others. As a matter of fact, *L. danicus* and *L. hargravesii,* the species with less divergence in morphology and specific genetic markers (Nanjappa et al., 2013), were found to be highly diverse regarding strain functions and microvariations based on sequences derived from the whole transcriptome. The ecological aspects of this kind of diversity could be seen in the spatial and temporal distribution; different species have different distribution, and also new clades identified in this study show signs of ecological adaptations to

different environments. The molecular differences detected in these clades were not many (at times single nucleotide differences), but they were found in V9, a small part of a scarcely variable phylogenetic marker and might correspond to wider variations in more variable markers. However, since there is no strong evidence on their characterization as new species, we currently refer to them as populations; as already stated in the Introduction, the present thesis treats groups of conspecific individuals that are demographically, genetically, or spatially disjunct from other groups of individuals, as populations. In the LTER-MC and more notably in the Tara Oceans dataset, populations with specific distributions were detected in *L. aporus* and *L. convexus*, the most abundant Leptocylindraceae species in Tara Oceans dataset (the population of *L. convexus* was not detected in LTER-MC dataset, where the species was less abundant than *L. danicus*). Both species showed populations slightly different in V9 from the already known ones, and in both cases these new populations showed a similar range and a preference for deep waters as well. In addition, the different population uncovered in *L. aporus* showed specific seasonal distribution too. Overall, this structure further supported the high intraspecific diversity in the most abundant species and pointed at ecological differences for those populations. The same was true for the entire species of *L. danicus* and *L. hargravesii,* with each one of them presenting different distribution patterns. Therefore, it is possible that the ecological separation of the populations mentioned above could even lead to speciation, if it has not already done so, as has been suggested for other eukaryotic (Palenik et al., 2007) and prokaryotic organisms (Kopac et al., 2014) and as might have happened in the case of *L. danicus* and *L. hargravesii*. Strain isolations in different seasons or in the regions where the new populations of *L. aporus*, *L. danicus* and *L. convexus* in Tara were detected should be performed in order to investigate if these are just populations or actual new species. In addition, multiple sampling at a single location across different seasons is proposed to verify the presence or absence of a species at a site.

Combining all the above information and the knowledge gained has allowed the diversity of the Leptocylindraceae family to be explored at many levels and ultimately identified functional and molecular differences among the species that could account for the seasonal and geographical

variation in their distributions. Consequently, it was possible to give some answers to the questions initially identified **(a)** on individual species and their composition (structure) across the season in terms of genetic and functional diversity and **(b)** on co-occurring species and their similar or distinct behaviors and responses to environmental conditions:

1. **Are individual species able to respond to environmental fluctuations due to populations consisting of similarly highly plastic individuals or rather populations consisting of diverse individuals each with narrower physiological tolerances and responses, which could also be a result of adaptation to different conditions?**

Plasticity and adaptation are terms that have been extensively described in the General Introduction, section 1.6. Differences in the genetic composition among populations in different regions/ seasons are considered to be a result of adaptation while in the case where each single individual in that population exhibits wide-ranging physiological capabilities, enabling it to cope with different conditions in different environments, is considered plasticity. In spite of the separate definitions given here, they are two mechanisms that cannot be easily distinguished since they are part of a dynamic process driven by environmental variability and natural selection. In the end, they both play an important role in diatoms distribution through their constant interactions. This was apparent in the current study, where signs of both plasticity, seen in physiological and functional responses of strains, and adaptation, seen in the distribution patterns of the new populations identified, were noticed. Highly plastic individuals of *Leptocylindrus* species might have been favored, especially in fluctuating environments, as they could rapidly acclimate to the new environmental conditions and then from there on, plasticity was altered by natural selection and ultimately became or facilitated adaptation. Therefore, to answer the question, it could be hypothesized that Leptocylindraceae, and possibly other diatoms, consist of populations with individuals of high plasticity (wider distribution) and populations of slightly narrower physiological tolerances and responses, which could also be a result of adaptation to different conditions. As it happens in most cases, the two scenarios do not have to exclude each other.

2. **Do different co-occurring species react in the same way under the same environmental conditions or might each one of them take advantage of different sections of the environmental spectrum and therefore have different functional traits?**

Just like in the individual species hypotheses, these two conditions are not mutually exclusive. Results of the present thesis revealed strategies, such as the TE stress response mechanism and the regulation of molecular transportation and signal transduction, which could be detected in most of the species, including co-occurring ones, while at the same time uncovered patterns that are confined to abundant species of broad seasonality or to rare ones with narrow seasonality separately.

Summing up, plasticity is a key characteristic of diatoms that allows them to survive in many different environments while it offers the appropriate phenotypic variability for selection to act on and lead to new adapted populations. If the specific environmental conditions reoccur and therefore the selection is strong, this could also lead to the evolution of new species. The high intraspecific plasticity, the constantly changing environment and the unconstrained access of diatoms to different niches in the sea could be the answer to the wide range of plankton species supported by a limited range of resources (plankton paradox); thus, the resources would be limited for plankton only in a specific time and place. The different models and hypotheses on the plankton paradox have been extensively discussed in the General Introduction, section 1.6, and after careful consideration of the current results, the case of Leptocylindraceae could be proposed to favor for the plankton diversity maintenance through non-equilibrium, as Hutchinson (1961) already suggested for all planktonic communities. Important mechanisms that take part in the complex processes of plasticity and adaptation and ultimately evolution of diatoms could include transposable elements that react to the environmental changes as has been shown in the case of *L. aporus* when faced with cold stress. In fact, temperature was proven to be an important factor determining the distribution and seasonality of the whole family. Although this thesis is based only on a single diatom family and contains a considerable degree of speculations, an explanation is suggested about the plankton paradox, an issue that has been already proposed

and explored in several studies (Hutchinson, 1961; Descamps-Julien et al., 2005; Miyazaki et al., 2006). In conclusion, diversity in Leptocylindraceae could be maintained because the habitat never reaches an equilibrium, ultimately leading to populations composed by highly plastic individuals but also populations with individuals which have adapted their plasticity in the cases of the more stable environments, like the deep sea and polar regions.

Overall, the present study adds valuable information in the context of phytoplankton research by pointing out that a re-assessment of a species diversity, using a larger dataset than used before (Nanjappa et al., 2014a) or more than one strain (growth and gene expression experiments), provides a more complete picture of the species ecology and evolution, which could have been totally missed previously due to the restriction of data or the exploration of variability on a single level. In addition, taking advantage of the new technologies to explore questions already investigated is a vital part of phytoplankton science, especially in genetics and transcriptomics. Similar interdisciplinary studies, combining physiological experiments, genetics and transcriptomics, and using more than one strain, in more diatom species could lead to the clarification and finally the generalization of concepts and theories applying to all diatoms regarding their seasonal and temporal distribution in the marine environment. However, the limitations that go along with each technique should be always kept in mind and considered when interpreting the results. For instance, one important drawback of transcriptomics in diatoms, which came up quite strongly in the present study, is the low percentage of annotation caused by the lack of extensive, well-curated datasets for free-living protists. The fact that many species-specific genes underlie the ecology of diatom species, and their interactions with the surrounding environment and co-occurring species (Bender et al., 2014; Pearson et al., 2015), makes their annotation even harder. As an indication of it, more than half of the proteins predicted in an iron limitation study in diatoms had no homology to known proteins (Mock et al., 2008) while in another study, a significant proportion of *P. tricornutum* genes (13%) were diatom-specific (Rastogi et al., in revision, cited in Tirichine et al., 2017) and therefore no functional information about the proteins encoded by them is available. Gene annotation must be improved

substantially if we are to gain a better mechanistic understanding of the cellular pathways and processes that are affected by environmental changes or vary across species. Furthermore, transcriptomes are not genomes. Transcriptomes recover only the genes that are expressed at the time of sampling instead of all the genetic potential of a species. This issue can be addressed to some extent by combining multiple transcriptomes from a strain that is cultured under different environmental conditions, in the same way it was done here for *L. aporus* at three different temperatures. Yet of course, gene expression does not immediately translate into protein function because of the potential for post-transcriptional and post-translational processing. Post-translational modifications, as well as genomic features that regulate transcription, such as promoters and methylation, are still not well understood (Veluchamy et al., 2013). For these and other reasons, a combination of transcriptomic, proteomic, metabolomic and gene manipulation studies of marine protists is necessary to validate any inferences made from studies of gene expression and this is the path that should be followed for Leptocylindraceae as well. On top of this and as already described further above, there are more hypotheses to be investigated in order to completely unravel the diversity, and thus ecology and evolution, of Leptocylindraceae. As technologies are improved and our knowledge and understanding of the marine ecosystem and its players expands and matures, we will hopefully be able to better explore and explain the way in which marine microorganisms survive, are distributed and thrive through space and time.

7. Bibliography

Adelfi, M.G., Borra, M., Sanges, R., Montresor, M., Fontana, A., Ferrante, M.I. 2014. Selection and validation of reference genes for qPCR analysis in the pennate diatoms *Pseudo-nitzschia multistriata* and *P. arenysensis*. Journal of Experimental Marine Biology and Ecology, Vol. 451, pp. 74–81.

Al-Kubaisi, K.H., Schwantes, H. O. 1981. Cytophotometrische Untersuchungen zum Generationswechsel autotropher and heterotropher siphonaler Organismen (Vaucheria sessilis and Saprolegnia ferax). Nova Hedwigia, Vol. 34, pp. 301- 316.

Allen, A.E., LaRoche, J., Maheswari, U., Lommer, M., Schauer, N., Lopez, P.J., Finazzi, G., Fernie, A.R., Bowler, C. 2008. Whole-cell response of the pennate diatom *Phaeodactylum tricornutum* to iron starvation. PNAS, Vol. 105, no.30, pp. 10438 – 10443.

Allshire, R. C., Javerzat, J. P., Redhead, N. J. and Cranston, G. 1994. Position effect variegation at fission yeast centromeres. Cell, Vol. 76, pp. 157–169.

Alpert, P., and Simms, E.L. 2002. The relative advantages of plasticity and fixity in different environments: when is it good for a plant to adjust? Evolutionary Ecology, Vol. 16, pp. 285–297.

Alverson, A.J. 2008. Molecular systematics and the diatom species. Protist, Vol. 159, pp. 339-353.

Amaral-Zettler, L.A., McCliment, E.A., Ducklow, H.W., Huse, S.M. 2009. A method for studying protistan diversity using massively parallel sequencing of V9 hypervariable regions of small-subunit ribosomal RNA genes. PLoS ONE Vol.4, Issue 7, e6372.

Amato, A., Kooistra, W.H.C.F., Levialdi Ghiron, J.H., Mann, D.G., Proschold, T., Montresor, M., 2007. Reproductive isolation among sympatric cryptic species in marine diatoms. Protist, Vol. 158, pp.193–207.

Andersen, C.L., Ledet-Jensen, J., Ørntoft, T. 2004. Normalization of real-time quantitative RT-PCR data: a model based variance estimation approach to identify genes suited for normalization - applied to bladder- and colon-cancer data-sets. Cancer Research, Vol. 64, pp. 5245-5250.

Andersen, R. 2005. Algal Culturing Techniques. Chapter 18: Measuring growth rates in microalgal cultures. Academic Press, USA.

Anderson, D.M., Kaefer, B.A., 1987. An endogenous annual clock in the toxic marine dinoflagellate Gonyaulax tamarensis. Nature, Vol. 325, pp. 616–617.

Ansari, K., Walter, S., Brennan, J.M., Lemmens, M., Kessans, S., McGahern, A., et al. 2007. Retrotransposon and gene activation in wheat in response to mycotoxigenic and non-mycotoxigenic-associated *Fusarium* stress. Theoretical and Applied Genetics, Vol. 114, pp. 927–937.

Anway, M.D., Cupp, A.S., Uzumcu, M., Skinner, M.K. 2005. Epigenetic transgenerational actions of endocrine disruptors and male fertility. Science, Vol. 308, pp. 1466–1469.

Armbrust, E. V., Berges, J.A., Bowler, C., Green, B.R., Martinez, D., et al. 2004. The genome of the diatom *Thalassiosira pseudonana*: ecology, evolution, and metabolism. Science, Vol. 306, pp. 79–86.

Armbrust, E.V. 2009. The life of diatoms in the world's oceans. Nature, Vol. 459, pp. 185–192.

Babraham Bioinformatics, FastQC. http://www.bioinformatics.babraham.ac.uk/projects/fastqc/

Bairoch, A. and Apweiler, R. 1996. The SWISS-PROT protein sequence data bank and its new supplement TREMBL. Nucleic Acid Research, Vol.24, Issue 1, pp. 21-25.

Barrick, J. E., Yu, D. S., Yoon, S. H., Jeong, H., Oh, T. K., et al. 2009. Genome evolution and adaptation in a longterm experiment with *E. coli*. Nature, Vol. 461, pp. 1243–1247.

Bayer-Giraldi, M., Uhlig, C., John, U., Mock, T., Valentin, K. 2010. Antifreeze proteins in polar sea ice diatoms: diversity and gene expression in the genus Fragilariopsis. Environmental Microbiology, Vol. 12, Issue 4, pp. 1041-52.

Beauchemin, M., Sougata, R., Pelletier, S., Averback, A., Lanthier, F., Morse, D. 2016. Characterization of two dinoflagellate cold shock domain proteins. Molecular Biology and Physiology, Vol. 1, Issue 1, e00034-15.

Becks, L., Ellner, S. P., Jones, L. E., Hairston, N.G. Jr. 2010. Reduction of genetic diversity radically alters eco-evolutionary community dynamics. Ecology Letters, Vol. 13, pp. 989–997.

Behnke, A., Friedl, T., Chepurnov, V.A., Mann, D. 2004. Reproductive compatibility and rDNA sequence analyses in the *Sellaphora pupula* species complex (Bacillariophyta). Journal of Phycology, Vol. 40, pp. 193-208.

Behnke, A., Engel, M., Christen, R., Nebel, M., Kleln, R. R. & Stoeck, T. 2011. Depicting more accurate pictures of protistan community complexity using pyrosequencing of hypervariable SSU rRNA gene regions. Environmental Microbiology, Vol. 13, pp. 340–349.

Bell, J.T., Pai, A.A., Pickrell, J.K., Gaffney, D.J., Pique-Regi, R., Degner, J.F., Gilad, Y., Pritchard, J.K. 2011. DNA methylation patterns associate with genetic and gene expression variation in HapMap cell lines. Genome Biology, Vol. 12, Issue 1, R10.

Bender, S.J., Durkin, C.A., Berthiaume, C.T., Morales, R.L., Armbrust, E.V. 2014. Transcriptional responses of three model diatoms to nitrate limitation of growth. Frontiers in Marine Science, Vol. 1, Article 3.

Berendsen, E.M., Koning, R.A., Boekhorst, J., de Jong, A., Kuipers, O.P. and Wells-Bennik, M.H.J. 2016. High-level heat resistance of spores of *Bacillus amyloliquefaciens* and *Bacillus licheniformis* results from the presence of a spoVA operon in a Tn1546 transposon. Frontiers in Microbiology, Vol. 7, 1912.

Berg, J., Appiano, M., Martínez, M.S., Hermans, F.W., Vriezen, W.H., Visser, R.G., et al. 2015. A transposable element insertion in the susceptibility gene *CsaMLO8* results in hypocotyl resistance to powdery mildew in cucumber. BMC Plant Biology, Vol. 15, 243.

Bhadury, P., Song, B., Ward, B.B. 2011. Intron features of key functional genes mediating nitrogen metabolism in marine phytoplankton. Marine Genomics, Vol. 4, pp. 207–213.

Bibby, T. S., Mary, I., Nield, J., Partensky, F. and Barber, J. 2003. Low-light-adapted *Prochlorococcus* species possess specific antennae for each photosystem. Nature, Vol. 424, pp. 1051–1054.

Bininda-Emonds, O. R. P., Gittleman, J. L. and Steel, M. A. 2002. The (super)tree of life: procedures, problems, and prospects. Annual Review of Ecology, Evolution and Systematics, Vol. 33, pp. 265–289.

Bolger, A. M., Lohse, M., and Usadel, B. 2014. Trimmomatic: A flexible trimmer for Illumina Sequence Data. Bioinformatics, Vol. 30, Issue 15, pp. 2114-20.

Bonduriansky, R. 2012. Rethinking heredity, again. Trends in Ecology and Evolution, Vol. 27, pp. 330–6.

Bouvet, G.F., Jacobi, V., Plourde, K.V. and Bernier, L. 2008. Stress-induced mobility of OPHIO1 and OPHIO2, DNA transposons of the Dutch elm disease fungi. Fungal Genetics and Biology, Vol. 45, pp. 565–578.

Bowen De Leon, K., Ramsay, B.D., Fields, M.W. 2012. Quality-score refinement of SSU rRNA gene pyrosequencing differs across gene region for environmental samples. Microbial Ecology, Vol. 64, pp. 499–508.

Bowler, C., Allen, A.E., Badger, J.H., Grimwood, J., Jabbari, K., Kuo, A., et al. 2008. The *Phaeodactylum* genome reveals the evolutionary history of diatom genomes. Nature, Vol. 456, pp. 239–244.

Boyd, P. W., Rynearson, T. A., Armstrong, E.A., Fu, F., Hayashi, K., et al. 2013. Marine Phytoplankton temperature versus growth responses from polar to tropical waters – outcome of a scientific community-wide study. PlosOne, Vol. 8, Issue 5, e63901.

Braarud, T. 1961. Cultivation of marine organisms as a means of understanding environmental influences on populations. In: Oceanography, M. Sears, ed., Publication No. 67, American Association of Advancement of Science, Washington, DC, pp. 271-298.

Bradford, B.U., Lock, E.F., Kosyk, O., Kim, S., Uehara, T., et al. 2011. Interstrain differences in the liver effects of trichloroethylene in a multistrain panel of inbred mice. Toxicological Sciences, Vol. 120, Issue 1, pp. 206-217.

Bradshaw, A. D. 1965. Evolutionary significance of phenotypic plasticity in plants. Advances In Genetics, Vol. 13, pp. 115–155.

Bray, J. R. and Curtis., J.T. 1957. An ordination of upland forest communities of southern Wisconsin. Ecological Monographs, Vol. 27, pp. 325-349.

Brewer A., Williamson M. 1994. A new relationship for rarefaction. Biodiversity and Conservation, Vol. 3, Issue 4, pp. 373–379.

Brion, C., Pflieger, D., Friedrich, A., Schacherer, J. 2015. Evolution of intraspecific transcriptomic landscapes in yeasts. Nucleic Acids Research, Vol. 43, No.9, pp. 4558-4568.

Brown, J. R., Douady, C. J., Italia, M. J., Marshall, W. E. and Stanhope., M. J. 2001. Universal trees based on large combined protein sequence data sets. Nature Genetics, Vol. 28, pp. 281–285.

Buckley, B.A., Burkhart, K.B., Gu, S.C., Spracklin, G., Kershner, A., et al. 2012. A nuclear Argonaute promotes multigenerational epigenetic inheritance and germline immortality. Nature, Vol. 489, pp. 447–51.

Buhmann, M.T., Poulsen, N., Klemm, J., Kennedy, M.R., Sherrill, C.D., Kroger, N. 2014. A tyrosine-rich cell surface protein in the diatom *Amphora coffeaeformis* identified through transcriptome analysis and genetic transformation. Plos One, Vol. 9, Issue 11, e110369.

Busch, W., Wunderlich, M., Schöffl, F. 2005. Identification of novel heat shock factor-dependent genes and biochemical pathways in *Arabidopsis thaliana*. The Plant Journal, Vol.41, pp. 1–14.

Cao, Y., Ohwatari, N., Matsumoto, T., Kosaka, M., Ohtsuru, A., Yamashita, S. 1999. TGF-β11 mediates 70-kDa heat shock protein induction due to ultraviolet irradiation in human skin fibroblasts. Pflügers Archiv, Vol. 438, Issue 3, pp. 239–244.

Caporaso, J.G., Kuczynski, J., Stombaugh, J, Bittinger K., Bushman, F.D., et al. 2010. QIIME allows analysis of high-throughput community sequencing data. Nature Methods, Vol. 7, pp. 335–336.

Caporaso, J.G., Paszkiewicz, K., Field, D., Knight, R., Gilbert, J.A. 2012. The Western English Channel contains a persistent microbial seed bank. ISME Journal, Vol.6, pp. 1089–1093.

Capy, P., Gasperi, G., Biemont, C. and Bazin, C. 2000. Stress and transposable elements: co-evolution or useful parasites? Heredity, Vol. 85, pp. 101–106.

Casteleyn, G., Leliaert, F., Backeljau, T., Debeer, A.-E., Kotaki, Y., et al. 2010. Limits to gene flow in a cosmopolitan marine planktonic diatom. Proceedings of the National Academy of Sciences, Vol. 107, pp. 12952–12957.

Castel, S.E., Martienssen, R.A. 2013. RNA interference in the nucleus: roles for small RNAs in transcription, epigenetics and beyond. Nature Review Genetics, Vol. 14, pp. 100–12.

Cavender-Bares, K.K., Karl, D.M. and Chisholm, S.W. 2001. Nutrient gradients in the western North Atlantic Ocean: Relationship to microbial community structure and comparison to patterns in the Pacific Ocean. Deep Sea Research Part I: Oceanographic Research Papers, Vol.48, pp. 2373-2395.

Cavrak, V., Lettner, N., Jamge, S., Kosarewicz, A., Bayer, L.M., Scheid, O.M. 2014. How a retrotransposon exploits the plant's heat stress response for its activation. PLOS Genetics, Vol.10, Issue 1, e1004115.

Chan, C. X., Reyes-Prieto, A., and Bhattacharya, D. 2011. Red and Green Algal Origin of Diatom Membrane Transporters: Insights into Environmental Adaptation and Cell Evolution. PLoS ONE, Vol. 6, Issue 12, e29138.

Chan, C.X., Bernard, G., Poirion, O., Hogan, J.M., Ragan, M.A. 2014. Interring phylogenies of evolving sequences without multiple sequence alignment. Scientific Reports, Vol. 4, Article n. 6504.

Chan, S.K., Hsing, M., Hormozdiari, F., Cherkasov, A. 2007. Relationship between insertion/deletion (indel) frequency of proteins and essentiality. BioMed Central Bioinformatics, Vol. 8, Issue 227.

Chandler, V.L., Stam, M. 2004. Chromatin conversations: mechanisms and implications of paramutation. Nature Review Genetics, Vol. 5, pp. 532–44.

Chattopadhyay, M.K., Raghu, G., Sharma, Y.V., Biju, A.R., Rajasekharan, M.V., Shivaji, S. 2011. Increase in oxidative stress at low temperature in an antarctic bacterium. Current Microbiology, Vol. 62, Issue 2, pp. 544-6.

Chenais, B., Caruso, A., Hiard, S., Casse, N. 2012. 2012. The impact of transposable elements on eukaryotic genomes: from genome size increase to genetic adaptation to stressful environments. Gene, Vol. 509, Issue 1, pp. 7-15.

Cheng, R., Feng, J., Zhang, B., Huang, Y., Cheng, J., Zhang, C. 2014. Transcriptome and gene expression analysis of an oleaginous diatom under different salinity conditions. Bioenergy Research, Vol. 7, pp. 192 -205.

Chepurnov, V.A., Mann, D.G., Sabbe, K. and Vyverman, W. 2004. Experimental studies on sexual reproduction in diatoms. International Review of Cytology, Vol. 237, pp. 91-154.

Chu, C.G., Tan, C.T., Yu, G.T., Zhong, S., Xu, S.S., Yan, L. 2011. A novel retrotransposon inserted in the Dominant Vrn-B1 allele confers spring growth habit in tetraploid wheat (*Triticum turgidum* L.). G3 (*Bethesda*, MD), Vol. 1, pp. 637–645.

Cingolani, P., Platts, A., Wang, L. L., Coon, M., Nguyen, T., Wang, L., Land, S, J., Lu, X., Ruden, D. M. 2012. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w^{1118}; iso-2; iso-3. Fly, Vol. 6, Issue 2, pp. 80–92.

Coate, J.E. and Doyle, J.J. 2015. Variation in transcriptome size: are we getting the message? Chromosoma, Vol. 124, Issue 1, pp. 27-43.

Colinet, H., Lee, S.F., Hoffmann, A. 2010. Temporal expression of heat shock genes during cold stress and recovery from chill coma in adult *Drosophila melanogaster*. FEBS Journal, Vol.277, Issue 1, pp.174-85.

Collins, S. 2011. Competition limits adaptation and productivity in a photosynthetic alga at elevated CO2. Proceedings of the Royal Society B: Biological Sciences, Vol. 278, Issue 1703, pp. 247–255.

Companion Web site: Figure W1 and Database W1 (available at http://Taraoceans.sb-roscoff.fr/EukDiv/).

Conesa, A., Madrigal, P., Tarazona, S., Gomez-Cabrero, D., Cervera, A., et al. 2016. A survey of best practices for RNA-seq data analysis. Genome Biology, Vol. 17, Issue 13.

Cooper, T., and Lenski., R. 2010. Experimental evolution with *E. coli* in diverse resource environments. I. Fluctuating environments promote divergence of replicate populations. BioMed Central Evolutionary Biology, Vol. 10, Issue 11.

Costas, E., Nieto, B., Lopez-Rodas, V., Salgado, C., Toro, M. 1998. Adaptation to competition by new mutation in clones of *Alexandrium minutum*. Evolution, Vol. 52, pp. 610–613.

Dai, J., Xie, W., Brady, T.L., Gao, J., Voytas, D.F. 2007. Phosphorylation regulates integration of the yeast Ty5 retrotransposon into heterochromatin. Molecular Cell, Vol. 27, pp. 289–299.

D'Alelio, D., Ribera d'Alcalà, M., Dubroca, L., Sarno, D., Zingone, A. and Montresor, M. 2010. The time for sex: A biennial life cycle in a marine planktonic diatom. American Society of Limnology and Oceanography, Vol. 55, Issue 1, pp.106-114.

Davison, I.R. 1991. Environmental effects on algal photosynthesis: temperature. Journal of Phycology, Vol. 27, pp. 2-8.

Decelle, J., Romac, S., Sasaki, E., Not, F., Mahe, F. 2014. Intracellular diversity of the V4 and V9 regions of the 18S rRNA in marine protists (radiolarians) assessed by high-throughput sequencing. PLoS One, Vol. 9, Issue 8, e104297.

Degerlund, M., Huseby, S., Zingone, A., Sarno, D., Landfald, B. 2012. Functional diversity in cryptic species of Chaetoceros socialis Lauder (Bacillariophyceae). Journal of Plankton Research, Vol.34, pp. 416–431.

de Jong, A., Van der Meulen, S., Kuipers, O.P. and Kok, J. 2015. T-Rex: Transcriptome analysis webserver for RNA-seq Expression data. BioMed Central Genomics, Vol.16, Issue 663.

Delaval, B., Bright, A., Lawson, N.D., Doxsey, S. 2011. The cilia protein IFT88 is required for spindle orientation in mitosis. Nature Cell Biology, Vol. 13, pp. 461 – 468

Delmont, T.O., Prestat, E., Keegan, K.P. et al. 2012. Structure, fluctuation and magnitude of a natural prairie soil metagenome. ISME Journal, Vol. 6, pp. 1677–1687.

Delsuc, F., Brinkmann, H., Philippe, H. 2005. Phylogenomics and the reconstruction of the tree of life. Nature Reviews Genetics, Vol. 6, Issue 5, pp. 361–75.

De Meester, L., Gomez, A., Okamura, B., Schwenk, K. 2002. The monopolization hypothesis and the dispersal-gene flow paradox in aquatic organisms. Acta Oecologica, Vol. 23, Issue 3, pp. 121–135.

Demott, W.R. and McKinney, E.N. 2015. Use it or lose it? Loss of grazing defenses during laboratory culture of the digestion-resistant green alga *Oocystis.* Journal of Plankton Research, Vol. 37, No. 2, pp. 1-10.

DeNicola D.M. 1996. Periphyton responses to temperature at different ecological levels, In: Stevenson R.J., Bothwell M.L., Lowe R.L. (Eds.), Algal Ecology: Freshwater Benthic Ecosystems, Academic Press, New York.

de Queiroz, Kevin. 2007. Species concepts and species delimitation. Society of Systematic Biologists, Vol. 56, Issue 6, pp. 879-886.

Derelle, R., Lopez-Garcia, P., Timpano, H. and Moreiara, D. 2016. A phylogenomic framework to study the diversity and evolution of Stramenopiles (=Heterokonts). Molecular Biology and Evolution, Online access: doi: 10.1093/molbev/msw168.

Descamps-Julien, B. and Gonzalez, A. 2005. Stable coexistence in a fluctuating environment: an experimental demonstration. Ecology, Vol. 86, Issue 10, pp. 2815 – 2824.

Deschamps, P. and Moreira, D. 2012. Reevaluating the green contribution to diatom genomes. Vol. 4, Issue 7, pp. 795-800.

de Vargas, C., Audic, S., Henry, N., Decelle, J., Mahé, F., et al. 2015. Eukaryotic plankton diversity in the sunlit ocean. Science, Vol. 348, pp. 1261605.

de Visser, J.A., Akkermans A.D., Hoekstra, R.F., de Vos, W.,M. 2004. Insertion-sequence-mediated mutations isolated during adaptation to growth and starvation in *Lactococcus lactis*. Genetics, Vol. 168, pp. 1145–1157.

Dewey, F. E., Pan, S., Wheeler, M. T., Quake, S. R., Ashley, E. A. 2012. DNA Sequencing: Clinical Applications of New DNA Sequencing Technologies. Circulation Journal Of The American Heart Association, Vol. 125, pp. 931-944.

Diaz-Real, J., Serrano, D., Piriz, A. and Jovani, R. 2015. NGS metabarcoding proves successful for quantitative assessment of symbiont abundance: the case of feather mites on birds. Experimental and Applied Acarology, Vol. 67, pp. 209-218.

Di Dato, V., Musacchia, F., Petrosino, G., Patil, S., Montresor, M., Sanges, R, Ferrante, M.I. 2015. Transcriptome sequencing of three *Pseudo-nitzschia* species reveals comparable gene sets

and the presence of nitrix oxide synthase genes in diatoms. Scientific Reports, Vol. 5, Article n. 12329.

Diersing, N. 2009. Phytoplankton blooms: the basics. National Oceanic and Atmospheric Administration, pp. 1-2.

Díez, B., Pedrós-Alió, C., and Massana, R. 2001. Study of genetic diversity of eukaryotic picoplankton in different oceanic regions by small-subunit rRNA gene cloning and sequencing. Applied and Environmental Microbiology, Vol. 67, pp. 2932-2941.

Dinsdale, E.A., Edwards, R.A., Hall, D., Angly, F., Breitbart,M. et al. 2008. Functional metagenomic profiling of nine biomes. Nature, Vol. 452, pp. 629–632.

Dobin, A., Davis, C.A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M.,Gingeras, T.R. 2012. STAR: ultrafast universal RNA-seq aligner. Bioinformatics, Vol. 29, Issue 1, pp. 15 - 21.

Downs C.A., McDougall, K.E., Woodley, C.M., Fauth, J.E., Richmond, R.H., et al. 2013. Heat-stress and light-stress induce different cellular pathologies in the symbiotic dinoflagellate during coral bleaching, PLoS ONE, Vol.8, Issue 12, e77173.

Drebes, G. 1977. Cell structure, cell division, and sexual reproduction of *Attheya decora* West (Bacillariophyceae, Biddulphiineae). Nova Hedwigia, Vol. 54, pp. 167-178.

Du, Z., Zhou, X., Ling, Y., Zhang, Z., and Su, Z. 2010. agriGO: a GO analysis toolkit for the agricultural community. Nucleic Acids Research Advance Access, Vol. 38, pp. W64-W70.

Dugar, G., Herbig, A., Förstner, K. U., Heidrich, N., Reinhardt, R., Nieselt, K., Sharma, C. M. 2013. High-resolution transcriptome maps reveal strain-specific regulatory features of multiple *Campylobacter jejuni* isolates. PLoS Genetics, Vol. 9, Issue 5, e1003495.

Duncan, E.J., Gluckman, P.D. and Dearden P.K. 2014. Epigenetics, plasticity and evolution: How do we link epigenetic change to phenotype? Journal of Experimental Zoology (Molecular and Developmental Evolution), Vol.322B, pp. 208-220.

Dunthorn, M., Klier, J., Bunge, J. and Stoeck, T. 2012. Comparing the hyper-variable V4 and V9 regions of the small subunit rDNA for assessment of ciliate environmental diversity. Journal of Eukaryotic Microbiology, Vol.59, Issue 2, pp. 185-187.

Dyhrman, S.T., Jenkins, B.D., Rynearson, T.A., Saito, M.A., Mercier, M.L., et al. 2012. The transcriptome and proteome of the diatom *Thalassiosira pseudonana* reveal a diverse phosphorus stress response. PLoS ONE, Vol. 7, Issue 3, e33768.

Edgar, R.C. 2013. UPARSE: Highly accurate OTU sequences from microbial amplicon reads. Nature Methods, Vol. 10, pp. 996–998.

Edgcomb, V., Orsi, W., Bunge, J., Jeon, S., Christen, R., Leslin, C. et al. 2011. Protistan microbial observatory in the Cariaco Basin, Caribbean. I. Pyrosequencing vs Sanger insights into species richness. The ISME Journal, Vol. 5, pp. 1344-1356.

Egue, F., Chenais, B., Tastard, E., Marchand, J., Hiard, S., Gateau, H., Hermann, D., Morant-Manceau, A., Casse, N., Caruso, A. 2015. Expression of the retrotransposons *Surcouf* and *Blackbeard* in the marine diatom *Phaeodactylum tricornutum* under thermal stress. Phycologia, Vol. 54, No. 6, pp. 617-627.

Eisen, J. A. and Fraser, C. M. 2003. Phylogenomics: intersection of evolution and genomics. Science, Vol. 300, pp. 1706–1707.

Elbrecht, V. and Leese, F. 2015. Can DNA-based ecosystem assessments quantify species abundance? Testing primer bias and biomass – sequence relationships with an innovative metabarcoding protocol. PLoS ONE, Vol. 10, Issue 7, e0130324.

Eppley, R.W. 1972. Temperature and phytoplankton growth in the sea. Fishery Bulletin, Vol. 70, pp. 1063–1085.

Everroad, R. C. and Wood., A.M. 2012. Phycoerythrin evolution and diversification of spectral phenotype in marine *Synechococcus* and related picocyanobacteria. Molecular Phylogenetics and Evolution, Vol. 64, pp. 381–392.

Falkowski, P.G., Raven, J.A. 1997. Aquatic photosynthesis. Blackwell Science, Oxford.

Falkowski, P.G., Barber, R.T., Smetacek, V. 1998. Biogeochemical controls and feedbacks on ocean primary production. Science, Vol. 281, pp.200-206.

Falkowski, P.G., Katz, M.E., Knoll, A.H., Quigg, A., Raven, J.A., Schofield, O., Taylor, F.J. 2004. The evolution of modern eukaryotic phytoplankton. Science, Vol. 305, Issue 5682, pp. 354-60.

Falkowski, P. G. and Knoll, A. H. 2007. Evolution of primary producers in the Sea. Elsevier Science Publishing, San Diego, US.

Feder, M.E., Hoffman, G.E. 1999. Heat-shock proteins, molecular chaperones, and the stress response: evolutionary and ecological physiology. Annual Review of Physiology, Vol.61, pp. 243–282

Felsenstein, J. 2004. Inferring Phylogenies. Sinauer Associates, Sunderland, Massachusetts.

Fenchel, T. and Finlay, B.J. 2004. The ubiquity of small species: patterns of local and global diversity, BioScience, Vol. 54, No. 8, pp. 777-784.

Feng, Y., Hare, C.E., Leblanc, K., Rose, J.M., Zhang, Y., et al. 2009. Effects of increased pCO2 and temperature on the North Atlantic spring bloom. I. The phytoplankton community and biogeochemical response. Marine Ecology - Progress Series, Vol. 388, pp. 13–25.

Feschotte, C. and Mouchès, C. 2000. Evidence that a family of miniature inverted-repeat transposable elements (MITEs) from the *Arabidopsis thaliana* genome has arisen from a pogo-like DNA transposon. Molecular Biology and Evolution, Vol. 17, pp. 730–737.

Field, C.B., Behrenfeld, M.J., Randerson, J.T., Falkowski, P. 1998. Primary production of the biosphere: integrating terrestrial and oceanic components. Sciene, Vol. 281, Issue 5374, pp. 237-240.

Figuerola, J., Green, A.J. 2002. Dispersal of aquatic organisms by waterbirds: a review of past research and priorities for future studies. Freshwater Biology, Vol. 47, Issue 3, pp. 483–494.

Flahaut, S., Benachour, A., Giard, J.C., Boutibonnes, P., Auffray, Y. 1996. Defense against lethal treatments and de novo protein synthesis induced by NaCl in *Enterococcus faecalis* ATCC 19433. Archives of Microbiology, Vol. 165, pp. 317–324.

Flick, K. and Kaiser, P. 2012. Protein degradation and the stress response. Seminars in Cell and Developmental Biology, Vol. 23, Issue 5, pp. 515-522.

Foissner, W. 2008. Protist diversity and distribution: some basic considerations, Biodiversity and Conservation, Vol. 17, No. 2, pp. 235-242.

Form, A. F., and Riebesell. U. 2012. Acclimation to ocean acidification during long-term CO2 exposure in the cold-water coral *Lophelia pertusa*. Global Change Biology, Vol.18, pp. 843–853.

Fourtanier, E. and Kociolek, J.P. 1999. Catalogue of the diatom genera. Diatom Research, Vol. 14, Issue 1, pp. 1-190.

Fu, F.X., Warner, M.E., Zhang, Y., Feng, Y., Hutchins, D.A. 2007. Effects of increased temperature and CO2 on photosynthesis, growth and elemental ratios of marine Synechococcus and Prochlorococcus (Cyanobacteria). Journal Phycology, Vol. 43, pp. 485–496.

Fusco, G. and Minelli, A. 2010. Phenotypic plasticity in development and evolution: facts and concepts. Introduction. Philosophical Transactions of the Royal Society of London B Biological Sciences, Vol.365, pp. 547–556.

Gaffè, J., McKenzie, C., Maharjan, R.P., Coursange, E., Ferenci, T., Schneider, D. 2011. Insertion sequence-driven evolution of *Escherichia coli* in chemostats. Journal of Molecular Evolution, Vol. 72, pp. 398–412.

Galinier, R., Roger, E., Monè, Y., Duval, D., Portet, A. 2017. A multistrain approach to studying the mechanisms underlying compatibility in the interaction between *Biomphalaria glabrata* and *Schistosoma mansoni.* PLoS Neglected Tropical Diseases, Vol. 11, Issue 3, e0005398.

Garber, M., Grabherr, M.G., Guttman, M. and Trapnell, C. 2011. Computational methods for transcriptome annotation and quantification using RNA-seq. Nature Methods, Vol. 8, No.6, pp. 469-477.

Garbuz, D.G. and Evgen'ev, M.B. 2017. The evolution of heat shock genes and expression patterns of heat shock proteins in the species from temperature contrasting habitats. Russian Journal of Genetics, Vol. 53, No. 1, pp. 21-38.

Geisen, S., Laros, I., Vizcaino, A., Bonkowski, M. and De Groot, G.A. 2015. Not all free-living: high-throughput DNA metabarcoding reveals a diverse community of protists parasitizing soil metazoan. Molecular Ecology, Vol.24, Issue 17, pp. 4556-4569.

Geitler, L. 1932. Der formwechsel der pennaten diatomeen (Kieselalgen). Archiv für Protistenkunde, Vol. 78, pp. 1-226.

Genitsaris, S., Monchy, S., Viscogliosi, E., Sime-Ngando, T., Ferreira, S., and Christaki, U. 2015. Seasonal variations of marine protist community structure based on taxon-specific traits using the eastern English Channel as a model coastal system. FEMS Microbiology Ecology , Vol. 91, pp. 1-15.

Georges, C., Monchy, S., Genitsaris, S., and Christaki, U. 2014. Protist community composition during early phytoplankton blooms in the naturally iron-fertilized Kerguelen area (Southern Ocean). Biogeosciences, Vol. 11, pp. 5847-5863.

Ghalambor, C.K., McKay, J.K., Carroll, S.P. and Reznick, D.N. 2007. Adaptive versus non-adaptive phenotypic plasticity and the potential for contemporary adaptation in new environments. Functional Ecology, Vol. 21, pp. 394 – 407.

Ghalambor, C.K., Hoke, K.L., Ruell, E.W., Fischer, E.K., Reznick D.N. and Hughes, K.A. 2015. Non-adaptive plasticity potentiates rapid adaptive evolution of gene expression in nature. Nature Research Letter, Vol. 525, pp. 372-375.

Giovannoni, S.J. and Vergin, K.L. 2012. Seasonality in ocean microbial communities. Science, Vol. 335, pp. 671 – 6.

Girard, L. and Freeling, M. 1999. Regulatory changes as a consequence of transposon insertion. Developmental Genetics, Vol. 25, pp.291–296.

Glatz, A., Vass, I., Los, D.A., Vigh, L. 1999. The Synechocystis model of stress: from molecular chaperones to membranes, Plant Physiology and Biochemistry Vol. 37, pp. 1-12

Glenn, T.C. 2011. Field guide to next-generation DNA sequencers. Molecular Ecology Resources, Vol. 11, Issue 5, pp. 759 – 769.

Gocheva, Y.G., Tosi, S., Krumova, E.T., Slokoska, L.S., Miteva, J.G., Vassilev Sv., Angelova, M.B. 2009. Temperature downshift induces antioxidant response in fungi isolated from Antarctica. Extremophiles, Vol. 13, pp.273–281.

Godhe, A., McQuoid, M.R., Karunasagar, I., Rehnstam-Holm, A.S. 2006. Comparison of three common molecular tools for distinguishing among geographically separated clones of the diatom *Skeletonema marinoi* Sarno et Zingone (Bacillariophyceae). Journal of Phycology, Vol. 42, Issue 2, pp. 280–291.

Godhe, A., Asplund, M.E., Härnström, K., Saravanan, V., Tyagi, A., Karunasagar, I. 2008. Quantification of diatom and dinoflagellate biomasses in coastal marine seawater samples by real-time PCR. Applied and Environmental Microbiology, Vol. 74, Issue 23, pp. 7174–7182.

Goldman, J., Ryther, J.H. 1976. Temperature-influenced species competition in mass cultures of marine phytoplankton. Biotechnology and Bioengineering, Vol. 18, Issue 8, pp.1125-1144.

Gonzàlez, J., Karasov, T.L., Messer, P.W., Petrov, D.A. 2010. Genome-wide patterns of adaptation to temperate environments associated with transposable elements in Drosophila. PLoS Genetics, Vol.6, e1000905.

Gotelli, N. J., Colwell, R. K. 2001. Quantifying biodiversity: procedures and pitfalls in the measurement and comparison of species richness. Ecology Letters, Vol. 4, pp. 379–391.

Gotthard, K. and Nylin, S. 1995. Adaptive plasticity and plasticity as an adaptation: a selective review of plasticity in animal morphology and life history. Oikos, Vol. 74, Issue 1, pp. 3-17.

Gouy, M., Guindon, S. and Gascuel, O. 2010. SeaView version 4: a multiplatform graphical user interface for sequence alignment and phylogenetic tree building. Molecular Biology and Evolution, Vol. 27, Issue 2, pp. 221-224.

Grabherr, M.G., Haas, B.J., Yassour, M., Levin, J.Z., Thompson, D.A., et al. 2011. Full-length transcriptome assembly from RNA-seq data without a reference genome. Nature Biotechnology, Vol.29, Issue 7, pp. 644-52.

Graham, L. E., Graham, J. M. and Wilcox, L. 2009. Algae, 2nd edition. Benjamin Cummings Pearson, San Francisco.

Grandbastien, M.A., Audeon, C., Bonnivard, E. et al. 2005. Stress activation and genomic impact of Tnt1 retrotransposons in Solanaceae. Cytogenetic and Genome Research, Vol. 110, pp. 229–241.

Griffiths-Jones, S., Bateman, A., Marshall, M., Khanna, A., Eddy, S.R. 2003. Rfam: an RNA family database. Nucleic Acids Research, Vol. 31, Issue 1, pp. 439-441.

Grzymski, J.J., Murray, A.E., Campbell, B.J., Kaplarevic, M., Gao, G.R., et al. 2008. Metagenome analysis of an extreme microbial symbiosis reveals eurythermal adaptation and metabolic flexibility. Proceedings of the National Academy of Sciences USA. Vol. 105, pp. 17516–17521.

Gsell, A. S., Domis, L. N. 2012. Genotype-by-temperature interactions may help to maintain clonal diversity in *Asterionella Formosa* (Bacillariophyceae). Phycological Society of America, Vol. 48, pp. 1197 – 1208.

Gu, Z., Rifkin, S.A., White, K.P., Li, W.H. 2004. Duplicate genes increase gene expression diversity within and between species. Nature Genetics, Vol. 36, pp. 577–579.

Guillard, R.L. and Ryther, J.H. 1962. Studies of marine planktonic diatoms. I. *Cyclotella nana Hustedt*, and *Detonula confervacea (Cleve) Gran*. Canadian Journal of Microbiology, Vol. 8, pp. 229-239.

Guillou, L., Bachar, D., Audic, S., Bass, D., Berney, C., et al. 2012. The Protist Ribosomal Reference database (PR2): a catalog of unicellular eukaryote Small Sub-Unit rRNA sequences with curated taxonomy. Nucleic Acids Research, Vol. 41, pp. 1-8.

Guindon, S., Dufayard, J.F., Lefort, V., Anisimova, M., Hordijk, W., Gascuel, O. 2010. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. Systematic Biology, Vol. 59, Issue 3, pp. 307-21.

Guo, R., Lee, M.A., Ki, J.S. 2013. Different transcriptional responses of heat shock protein 70/90 in the marine diatom *Ditylum brightwellii* exposed to metal compounds and endocrine-disrupting chemicals. Chemosphere, Vol. 92, Issue 5, pp.535-43.

Guo, Y., Levin, H.L. 2010. High-throughput sequencing of retrotransposon integration provides a saturated profile of target activity in *Schizosaccharomyces pombe*. Genome Research, Vol. 20, pp. 239–248.

Gutierrez-Arcelus, M., Lappalainen, T., Montgomery, S.B., Buil, A., Ongen, H., et al. 2013. Passive and active DNA methylation and the interplay with genetic variation in gene regulation. eLife, Vol. 2, e00523.

Haag, C.R., Riek, M., Hottinger, J.W., Pajunen, V.I., Ebert, D. 2006. Founder events as determinants of within-island and among-island genetic structure of Daphnia metapopulations. Heredity, Vol. 96, Issue 2, pp. 150–158.

Hadziavdic, K., Lekang, K. Lanzen, A., Johassen, I. Thompson, E.M. and Troedsson, C. 2014. Characterization of the 18s rRNA gene for designing universal eukaryote specific primers. PlosOne, Vol.9, Issue 2, e87624.

Hall, T.A. 1999. BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. Nucleic Acids Symposium Series. Vol. 41, pp. 95–98.

Hamblin, M.T., Di Rienzo, A. 2000. Detection of the signature of natural selection in humans: evidence from the Duffy Blood group locus. The American Journal of Human Genetics, Vol. 66, pp. 1669–1679.

Hansen, S. R., and Hubbell, S.P. 1980. Single-nutrient microbial competition: Agreement between experimental and theoretically forecast outcomes. Science, Vol. 207, pp. 1491-1493.

Hare, C., Leblanc, K., DiTullio, G.R., Kudela, K.M., Zhang, Y., et al. 2007. Consequences of increased temperature and CO2 for phytoplankton community structure in the Bering Sea. Marine Ecology - Progress Series, Vol. 352, pp. 9–16.

Hartl, F. 1996. Molecular chaperones in cellular protein folding. Nature, Vol.381, pp. 571–579.

Hasle, G.R., Syvertsen, E.E. 1997. Marine diatoms, Identifying marine phytoplankton, Academic Press: San Diego, pp. 5-385.

Heidelberg, K. B., Gilbert, J. A. and Joint, I. 2010. Marine genomics: at the interface of marine microbial ecology and biodiscovery, Microbial Biotechnology, Vol. 3, Issue 5, pp. 531-543.

Helbling, E.W., Buma, A.G.J., Boelen, P., van der Strate, H., Giordanino, M.V.F., Villafane, E. 2011. Increase in Rubisco activity and gene expression due to elevated temperature partially counteracts ultraviolet radiation–induced photoinhibition in the marine diatom *Thalassiosira weissflogii.* Limnology and Oceanography, Vol.56, Issue 4, pp. 1330-1342.

Henderson, R. J., E. N. Hegseth, and M. T. Park. 1998. Seasonal variation in lipid and fatty acid composition of ice algae from the Barents Sea. Polar Biology, Vol. 20, pp. 48-55.

Hendry, A.P., Day, T., 2005. Population structure attributable to reproductive time: isolation by time and adaptation by time. Molecular Ecology, Vol. 14, pp. 901–916.

Hermann, D., Egue, F., Tastard, E., Nguyen, D., Casse, N., et al. 2014. An introduction to the vast world of transposable elements – what about the diatoms? Diatom Research, Vol. 29, Issue 1, pp.91-104.

Hillebrand, H., Dürselen, C., Kirschtel, D., Pollingher, U. and Zohary, T. 1999. Biovolume calculation for pelagic and benthic microalgae. Journal of Phycology, Vol.35, pp. 403-424.

Höhl, M., Rigoutsos, I., Ragan, M.A. 2006. Pattern-based phylogenetic distance estimation and tree reconstruction. Evolutionary Bioinformatics Online, Vol. 2, pp. 359-375.

Höhl, M. and Ragan, M.A. 2007. Is multiple-sequence alignment required for accurate inference of phylogeny? Systematic Biology, Vol. 56, Issue 2, pp. 206-221.

Hormozdiari, F., Hsing, M., Salari, R., Schönhuth, A., Chan, S.K., Sahinalp, S.C., Cherkasov, A. 2009. Effect of insertions and deletions (indels) on wirings in protein-protein interaction networks: a large-scale study. Journal of Computational Biology, Vol.16, pp. 159–167.

Horn, G., Hofweber, R., Kremer, W., Kalbitzer, HR. 2007. Structure and function of bacterial cold shock proteins. Cellular and Molecular Life Sciences, Vol. 64, Issue 12, pp. 1457-70.

Hossain, M.M., Nakamoto, H. 2003. Role for the cyanobacterial HtpG in protection from oxidative stress. Current Microbiology, Vol.46, pp. 70–76.

Houle, D. 1992. Comparing evolvability and variability of quantitative traits. Genetics, Vol. 130, pp. 195–204.

Houston, A. I. and McNamara, J. M. 1992. Phenotypic plasticity as a state-dependent life-history decision. Evolutionary Ecology, Vol. 6, pp. 243-253.

Howe, E. A., Sinha, R., Schlauch, D., and Quackenbush, J. 2011. RNA-Seq Analysis in MeV. Bioinformatics, Vol. 22, pp. 3209-3210.

Hsu, S., Hubbell, S. and Waltman, P. 1977. A mathematical theory for single nutrient competition in continuous cultures of micro-organisms. Society for Industrial and Applied Mathematics, Vol. 32, pp. 366-383.

Huertas, I. E., Rouco, M., Lopez-Rodas, V. and Costas, E. 2010. Estimating the capability of different phytoplankton groups to adapt to contamination: herbicides will affect phytoplankton species differently. New Phytologist, Vol. 188, pp. 478–487.

Hulburt, E.M. and Guillard, R.R.L. 1968. The relationship of the distribution of the diatom *Skeletonema tropicum* to temperature. Ecology, Vol.49, pp. 337-9.

Hulburt, E.M. 1982. The adaptation of marine phytoplankton species to nutrient and temperature. Ocean Science Engineering, Vol.7, pp. 187-228.

Huse, S.M., Welch, D.M., Morrison, H.G., and Sogin, M.L. 2010. Ironing out the wrinkles in the rare biosphere through improved OTU clustering. Environmental Microbiology, Vol. 12, pp. 1889-1898.

Huseby, S., Degerlund, M., Zingone, A., Hansen, E., 2012. Metabolic fingerprinting reveals differences between northern and southern strains of the cryptic diatom *Chaetoceros socialis*. European Journal of Phycology, Vol. 47, pp.480–489.

Huson, D.,H., Auch, A.F., Schuster, S.C. 2007. MEGAN analysis of metagenomic data. Genome Research, Vol. 17, Issue 3, pp. 377-386.

Hutchinson, G.E. 1961. The paradox of the plankton. The American Naturalist, Vol.95, No. 882, pp.137-145.

Iglesias-Rodriguez, D., Schofield, O. M., Batley, J., Medlin, L.K. and Hayes, P.K. 2006. Intraspecific genetic diversity in the marine cocolithophore *Emiliania huxleyi* (Prymnesiophyceae): the use of microsatellite analysis in marine phytoplankton population studies. Journal of Phycology, Vol. 42, pp. 526–536.

Ito, H., Gaubert, H., Bucher, E., Mirouze, M., Vaillant, I., Paszkowski, J., 2011. An siRNA pathway prevents transgenerational retrotransposition in plants subjected to stress. Nature, Vol. 472, 115e119.

Jablonska, E, Raz, G. 2009. Transgenerational epigenetic inheritance: prevalence, mechanisms, and implications for the study of heredity and evolution. The Quarterly Review of Biology, Vol. 84, pp. 131–76.

Janska, A., Aprile, A., Cattivelli, L., Zámečníket, J., de Bellis, L., Ovesnà, J. 2014. The up-regulation of elongation factors in the barley leaf and the down-regulation of nucleosome assembly genes in the crown are both associated with the expression of frost tolerance. Functional and Integrative Genomics, Vol. 14, Issue, pp. 493–506.

Jeffroy, O., Brinkmann, H., Delsuc, F., Philippe, H. 2006. Phylogenomics: the beginning of incongruence?. Trends in Genetics, Vol. 22, Issue 4, pp. 225–31.

Jen-Pan Huang, 2015. Revisiting rapid phenotypic evolution in sticklebacks: integrative thinking of standing genetic variation and phenotypic plasticity. Frontiers in Ecology and Evolution, Vol. 3, Article 47.

Jia, S., Noma, K. and Grewal, S. I. 2004. RNAi-independent heterochromatin nucleation by the stress-activated ATF/CREB family proteins. Science, Vol. 304, pp. 1971–1976.

Jin, P., Gao, K. and Beardall, J. 2013. Evolutionary responses of a coccolithophorid *Geophyrocapsa oceanica* to ocean acidification. Evolutionary Biology, Vol. 67, pp. 1869–1878.

Junemann, S., Sedlazeck, F. J., Prior, K., Albersmeier, A., John, U., et al. 2013. Updating benchtop sequencing performance comparison. Nature Biotechnology, Vol. 31, pp. 294–296.

Jung, G., Lee, C.G., Kang, S.H., Jin, E. 2007. Annotation and expression profile analysis of cDNAs from Antarctic diatom *Chaetoceros neogracile*. Journal of Microbiology and Biotechnology, Vol. 17, Issue 8, pp. 1330-7.

Kampinga, H.H., Brunsting, J.F., Stege, G.J., Burgman, P.W., Konings, A.W. 1995. Thermal protein denaturation and protein aggregation in cells made thermotolerant by various chemicals: role of heat shock proteins. Experimental Cell Research, Vol. 219, pp. 536–546.

Kanazawa, A., Liu, B., Kong, F., Arase, S., Abe, J. 2009. Adaptive evolution involving gene duplication and insertion of a novel Ty1/copia-like retrotransposon in soybean. Journal of Molecular Evolution, Vol. 69, pp. 164–175.

Kapoor, M., Sreenivasan, G.M., Goel, N., Lewis, J. 1990. Development of thermotolerance in *Neurospora crassa* by heat shock and other stresses eliciting peroxidase induction. Journal of Bacteriology, Vol. 172, pp. 2798–2801.

Karentz, D. and Smayda, T.J. 1984. Temperature and seasonal occurrence patterns of 30 dominant phytoplankton species in Narragansett Bay over a 22-year period (1959 – 1980). Marine Ecology – Progress Series, Vol. 18, pp.277-293.

Karlson, D., Imai, R. 2003. Conservation of the cold shock domain protein family in plants. Plant Physiology, Vol. 131, pp. 12–15.

Katoh, K. and Standley, D.M. 2013. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Molecular Biology and Evolution, Vol. 30, Issue 4, pp. 772-780.

Kauserud, H., Kumar, S., Brysting, A.K., Norden, J., Carlsen, T. 2012. High consistency between replicate 454 pyrosequencing analyses of ectomycorrhizal lant root samples. Mycorrhiza, Vol. 22, pp. 309–315.

Kawecki, T. J. and Stearns, S. C. 1993. The evolution of life histories in spatially heterogeneous environments: optimal reaction norms revisited. Evolutionary Ecology, Vol. 7, pp. 155-174.

Keeling, P.J., Burki, F., Wilcox, H.M., Allam, B., Allen, E.E., et al. 2014. The marine microbial eukaryote transcriptome sequencing project (MMETSP): Illuminating the functional diversity of eukaryotic life in the oceans through transcriptome sequencing. PLOS Biology, Vol. 12, Issue 6, e1001889.

Keller,M.D., Selvin, R.C., Claus,W. and Guillard, R.R.L. 1987. Media for the culture of oceanic ultraphytoplankton. Journal of Phycology, Vol. 23, pp. 633-638.

Kent, W.J., Sugnet, C.W., Furey, T.S., Roskin, K.M., Pringle, T.H., Zahler, A.M., Haussler, D. 2002. The human genome browser at UCSC. Genome Research, Vol.6, pp. 996-1006.

Kerr, B., Riley, M., Feldman, M., Bohannan, B. 2002. Local dispersal promotes biodiversity in a real-life game of rock–paper–scissors. Nature, Vol. 418, pp. 171–174.

Kidwell, M. G. 1977. Reciprocal differences in female recombination associated with hybrid dysgenesis in *Drosophila melanogaster*. Genetic Research, Vol. 30, pp. 77–88.

Kidwell, M. G. and Lisch, D. R. 2000. Transposable elements and host genome evolution. Trends in Ecology and Evolution, Vol. 15, pp. 95–99.

Kilham, P. and Kilham, S.S. 1980. The evolutionary ecology of phytoplankton. In I. Morris, ed. The Physiological Ecology of Phytoplankton. University of California Press, Berkeley.

Kim, B.H. and Schöffl, F. 2002. Interaction between *Arabidopsis* heat shock transcription factor 1 and 70 kDa heat shock proteins. Journal of Experimental Botany, Vol.53, pp. 371-375.

Kim, J., Nueda, A., Meng, Y.H., Dynan, W.S., Mivechi, N.F. 1997. Analysis of the phosphorylation of human heat shock transcription factor-1 by MAP kinase family members. Journal of Cellular Biochemistry, Vol. 67, pp.43–54.

Kim, K.M., Park, J., Bhattacharya, D. and Yoon, H.S. 2014. Applications of next-generation sequencing to unravelling the evolutionary history of algae. International Journal of Systematic and Evolutionary Microbiology, Vol. 64, pp. 333-345.

Kim, M.H., Sasaki, K., Imai, R. 2009. Cold shock domain protein 3 regulates freezing tolerance in *Arabidopsis thaliana.* Journal of Biological Chemistry, Vol. 284, pp. 23454–23460.

Kim, R., Guo, J.T. 2010. Systematic analysis of short internal indels and their impact on protein folding. BioMed Central Structural Biology, Vol.4, pp. 10–24.

Kim, W., Kim, H. D., Jung, Y., Kim, J., and Chung, J. 2015. *Drosophila* Low Temperature Viability Protein 1 (LTV1) is required for ribosome biogenesis and cell growth downstream of *Drosophila* Myc (dMyc). The Journal of Biological Chemistry, Vol. 290, Issue 21, pp. 13591–13604.

Kinoshita, S., Kaneko, G., Lee, J.H., Kikuchi, K., Yamada, H., Hara, T., Itoh, Y., Watabe, S. 2001. A novel heat stress-responsive gene in the marine diatom *Chaetoceros compressum* encoding two types of transcripts, a trypsin-like protease and its related protein, by alternative RNA splicing. European Journal of Biochemistry, Vol. 268, Issue 17, pp. 4599-609.

Klironomos, F., Berg, J., Collins, S. 2013. How epigenetic mutations can affect genetic evolution: Model and mechanism. Bioessays, Vol. 35, Issue 6, pp. 571-8.

Kneitel, J.M. and Chase, J.M. 2004. Trade-offs in community ecology: linking spatial scales and species coexistence. Ecology Letters, Vol. 7, pp. 69-80.

Koester, J.A., Swanson, W.J. and Armbrust, V.E. 2013. Positive selection within a diatom species acts on putative protein interactions and transcriptional regulation. Molecular Biology and Evolution, Vol. 30, Issue 2, pp. 422-434.

Kooistra, W.H.C.F., Gersonde, R., Medlin, L.K., Mann, D.G. 2007. The origin and evolution of the diatoms: their adaptation to a planktonic existence. In Evolution of Primary Producers in the Sea. Edited by Falkowski, P.G., Knoll AH. New York: Academic Press, pp. 207-249.

Kooistra, W.H.C.F., Sarno, D., Balzano, S., Gu, H., Andersen, R.A. and Zingone, A. 2008. Global diversity and biogeography of Skeletonema species (Bacillariophyta). Protist, Vol. 159, pp. 177-193.

Kooistra, W.H.C.F., Sarno, D., Hernández-Becerril, D.U., Assmy, P., Di Prisco, C. and Montresor, M. 2010. Comparative molecular and morphological phylogenetic analyses of taxa in the Chaetocerotaceae (Bacillariophyta). Phycologia, Vol. 49, pp. 471-500.

Kopac, S., Wang, Z., Wiedenbeck, J., Sherry, J., Wu, M., & Cohan, F. M. 2014. Genomic heterogeneity and ecological speciation within one subspecies of *Bacillus subtilis*. Applied and Environmental Microbiology, Vol. 80, Issue 16, pp. 4842–4853.

Kopf, A., Bicak, M., Kottmann, R., Schnetzer, J., Kostadinov, I., Lehmann, K. et al. 2015. The ocean sampling day consortium. GigaScience, Vol. 4, Issue 27.

Kozich, J.J., Westcott, S.L., Baxter, N.T., Highlander, S.K., and Schloss, P.D. 2013. Development of a dual-index sequencing strategy and curation pipeline for analyzing amplicon sequence data on the MiSeq Illumina sequencing platform. Applied and Environmental Microbiology, Vol. 79, pp. 5112-5120.

Kremp, A., Godhe, A., Egardt, J., Dupont, S., Suikkanen, S., Casabianca and Penna, A. 2012. Intraspecific variability in the response of bloom-forming marine microalgae to changed climate conditions. Ecology and Evolution, Vol.2, Issue 6, pp. 1195-1207.

Kristensen, D.M., Wolf, Y.I., Mushegian, A.R., Koonin, E.V. 2011. Computational methods for gene orthology inference. Briefings in Bioinformatics, Vol. 12, pp. 379–91.

Kristiansson, E., Österlund, T., Gunnarsson, L., Anre, G., Larsson, D.G.J. and Nerman, O. 2013. A novel method for cross-species gene expression analysis. BioMed Central Bioinformatics, Vol.14, Issue 70.

Kucharski, R., Maleszka, J., Foret, S., Maleszka, R. 2008. Nutritional control of reproductive status in honeybees via DNA methylation. Science, Vol.319, pp. 1827–1830.

Kumar, A., Bennetzen, J.L. 1999. Plant retrotransposons. Annual Review of Genetics, Vol.33, pp. 479–532.

Kunin, V., Engelbrektson, A., Ochman, H., Hugenholtz, P. 2010. Wrinkles in the rare biosphere: pyrosequencing errors can lead to artificial inflation of diversity estimates. Environmental Microbiology, Vol. 12, pp. 118–123.

Langmead, B., Salzberg, S. 2012. Fast gapped-read alignment with Bowtie 2. Nature Methods, Vol. 9, pp. 357-359.

Lakeman, M.B. and Cattolico, R. A. 2007. Cryptic diversity in phytoplankton cultures is revealed using a simple plating technique. Journal of Phycology, Vol. 43, Issue 4, pp. 662–674.

Lakeman, M.B., von Dassow, P. and Cattolico, R.A. 2009. The strain concept in phytoplankton ecology. Harmful Algae, Vol. 8, pp. 746-758.

Lander, E.S. et al. 2001. Initial sequencing and analysis of the human genome. Nature, Vol.409, pp.860-921.

Larkindale, J., Knight, M.R. 2002. Protection against heat stress-induced oxidative damage in *Arabidopsis* involves calcium, abscisic acid, ethylene, and salicylic acid. Plant Physiology, Vol. 128, pp. 682–695.

Larkindale, J., Mishkind, M., Vierling, E. 2007. Plant responses to high temperature. In MA Jenks, PM Hasegawa, eds, Plant Abiotic Stress. Blackwell Scientific Publications, Oxford.

Lassmann, T., Hayashizaki, Y., and Daub, C. O. 2011. SAMStat: monitoring biases in next generation sequencing data. Bioinformatics, Vol. 27, Issue 1, pp. 130–131.

Lauritano, C., Orefice, I., Procaccini, G., Romano, G. And Ianora, A. 2015. Key genes as stress indicators in the ubiquitous diatom *Skeletonema marinoi.* BioMed Central Genomics, Vol. 16, pp. 411.

Lee, B.H., Lee, H., Xiong, L., Zhu, J.K. 2002. A mitochondrial complex I defect impairs cold-regulated nuclear gene expression, Plant Cell, Vol. 14, pp. 1235–1251.

Levitan, O., Dinamarca, J., Zelzion, E., Lun, D.S., Guerra, L.T., Kim, M.K., Kim, J., Van Mooy, B.A., Bhattacharya, D., Falkowski, P.G. 2015. Remodeling of intermediate metabolism in the diatom *Phaeodactylum tricornutum* under nitrogen stress. Proceedings of the National Academy of Sciences of U. S. A., Vol. 112, pp. 412–417.

Li, B. and Dewey, C.N. 2011. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. BioMed Central Bioinformatics, Vol.12, Issue 323.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R. and 1000 Genome Project Data Processing Subgroup. 2009. The Sequence alignment/map (SAM) format and SAMtools. Bioinformatics, vol.25, pp. 2078-9.

Li, J., Huang, Q., Sun, M., Zhang, T., Li, H., et al. 2016. Global DNA methylation variations after short-term heat shock treatment in cultured microspores of *Brassica napus* cv. Topas. Science Reports, Vol. 6, 38401.

Li, W., Jaroszewski L. and Godzik, A. 2001. Clustering of highly homologous sequences to reduce the size of large protein database. Bioinformatics, Vol. 17, pp.282-283.

Lim, A.L., Ng, S., Leow, S.C., Choo, R., Ito, M. et al. 2012. Epigenetic state and expression of imprinted genes in umbilical cord correlates with growth parameters in human pregnancy. Journal of Medical Genetics, Vol. 49, pp. 689–697.

Lin, J. and Gerstein, M. 2000. Whole-genome trees based on the occurrence of folds and orthologs: implications for comparing genomes on different levels. Genome Research, Vol. 10, pp. 808–818.

Lindquist, S., 1986. The heat-shock response. Annual Review of Biochemistry, Vol. 55, pp. 1151–1191.

Liu, S., Graham, J.E., Bigelow, L., Morse, P.D. II, Wilkinson, B.J. 2002. Identification of *Listeria monocytogenes* genes expressed in response to growth at low temperature. Applied and Environmental Microbiology, Vol.68, pp. 1697–1705.

Liu, Y., Zhang, C., Chen, J., Guo, L., Li, X., Li, W., Yu, Z., Deng, J., Zhang, P., Zhang, K., Zhang, L. 2013. *Arabidopsis* heat shock factor HsfA1a directly senses heat stress, pH changes, and hydrogen peroxide via the engagement of redox state. Plant Physiology and Biochemistry, Vol. 64, p.92-98.

Llinás, M., Bozdech, Z., Wong, E.D., Adai, A.T., DeRisi, J.L. 2006. Comparative whole genome transcriptome of three *Plasmodium falciparum* strains. Nucleic Acids Research, Vol. 34, No. 4, pp. 1166 – 1173.

Loar, J. W., R. M., Seiser, A. E., Sundberg, H. J., Sagerson, N., Ilias, et al. 2004. Genetic and biochemical interactions among Yar1, Ltv1 and Rps3 define novel links between environmental stress and ribosome biogenesis in *Saccharomyces cerevisiae*. Genetics, Vol. 168, pp. 1877–1889.

Logares, R., Audic, S., Bass, D., Bittner, L., Boutte, C., Christen, R. et al. 2014. Patterns of rare and abundant marine microbial eukaryotes. Current Biology, Vol. 24, pp. 813-821.

Lohbeck, K. T., Riebesell, U. and Reusch, T.B.H. 2012. Adaptive evolution of a key phytoplankton species to ocean acidification. Nature Geoscience, Vol. 5, pp. 917–917.

Lomas, M. W., and P. M. Gilbert. 1999. Temperature regulation of nitrate uptake: a novel hypothesis about nitrate uptake and reduction in cool-water diatoms. Limnology and Oceanography, Vol. 44, pp.556–572.

Lommer, M., Specht, M., Roy, A., Kraemer, L., Anderson, R., et al. 2012. Genome and low-iron response of an oceanic diatom adapted to chronic iron limitation. Genome Biology, Vol. 13, R66.

Long, L., Ou, X., Liu, J., Lin, X., Sheng, L. and Liu, B. 2009. The spaceflight environment can induce transpositional activation of multiple endogenous transposable elements in a genotype-dependent manner in rice. Journal of Plant Physiology, Vol. 166, pp. 2035–2045.

Lopes, F., Jjingo, D., daSilva, C.R., Andrade, A.C., Marraccini, P., Teixeira, J.B., et al. 2013. Transcriptional activity, chromosomal distribution and expression effects of transposable elements in *Coffea* genomes. PLoS ONE, Vol. 8, e78931.

Lopez-Maestre, H., Brinza, L., Marchet, C., Kielbassa, J., Bastien, S., et al. 2016. SNP calling from RNA-seq data without a reference genome: identification, quantification, differential analysis and impact on the protein sequence. Nucleic Acids Research, Vol. 44, Issue 19, e148.

López-Rodas, V. L., Marva, F., Rouco, M., Costas, E. and Flores- Moya, A. 2008. Adaptation of the chlorophycean *Disctyosphaerium chlorelloides* to stressful acidic, mine metal-rich waters as a result of pre-selective mutations. Chemosphere, Vol. 72, pp. 703–707.

Loppes, R., N. Devos, S. Willem, P. Barthelemy, and R. Matagne. 1996. Effect of temperature on two enzymes from a psychrophilic Chloromonas (CHLOROPHYTA). Journal of Phycology, Vol. 32, pp.276–278.

Lynch, M., Gabriel, W. and Wood, A. M. 1991. Adaptive and demographic responses of plankton populations to environmental change. Limnology and Oceanography, Vol. 36, pp. 1301–1312.

Lundholm, N., Bates, S.S., Baugh, K.A., Bill, B.D., Connell, L.B., Léger, C. and Trainer, V.L. 2012. Cryptic and pseudo-cryptic diversity in diatoms—with descriptions of Pseudo-nitzschia hasleana sp. nov. and P. fryxelliana sp. nov. Journal of Phycology, Vol. 48, pp. 436-454.

Lutz, U., Posè, D., Pfeifer, M., Gundlach, H., Hagmann, J., et al. 2015. Modulation of ambient temperature-dependent flowering in *Arabidopsis thaliana* by natural variation of *FLOWERING LOCUS M*, PLoS Genetics, Vol. 11, Issue 10, e1005588.

Mahè, F., Rognes, T., Quince, C., de Vargas, C. and Dunthorn, M. 2014. Swarm: robust and fast clustering method for amplicon-based studies. PeerJ2, Vol. 2, e593.

Maheswari, U., Mock, T., Armbrust, V. and Bowler, C. 2009. Update of the Diatom EST Database: a new tool for digital transcriptomics. Nucleic Acids Research, Vol. 37, D1001–D1005.

Makarevitch, I., Waters, A.J., West, P.T., Stitzer, M., Hirsch, C.N., Ross-Ibara, J., Springer, N.M. 2015. Transposable elements contribute to activation of maize genes in response to abiotic stress. PLOS Genetics, Vol.11, Issue 1, e1004915.

Malviya, S., Scalco, E., Audic, S., Vincent, F., Veluchamy, A., Poulain, J. et al. 2016. Insights into global diatom distribution and diversity in the world's ocean. Proceedings of the National Academy of Sciences, Vol. 113, Issue 11, E1516-25.

Manikkam, M., Tracey, R., Guerrero-Bosagna, C., Skinner, M,K. 2012. Dioxin (TCDD) induces epigenetic transgenerational inheritance of adult onset disease and sperm epimutations. PLoS ONE, Vol. 7, e46249.

Mann, D. G. and Vanormelingen, P. 2013. An inordinate fondness? The number, distributions, and origins of diatom species. Journal of Eukaryotic Microbiology, Vol. 60, pp. 414-420.

Marchant, A., Mouqel, F., Mendonca, V., Quartier, M., Jacquin-Joly, E., da Rosa, J.A., Petit, E., Harry, M. 2016. Comparing de novo and reference-based transcriptome assembly strategies by applying them to the blood-sucking bug *Rhodnius prolixus*. Insect Biochemistry and Molecular Biology, Vol. 69, pp. 25-33.

Marchler-Bauer, A., Derbyshire, M.K., Gonzales, N.R., Lu, S., Chitsaz, F., Geer, L.Y., Geer, R.C., He, J., Gwadz, M., Hurwitz, D.I., Lanczycki, C.J., Lu, F., Marchler, G.H., Song, J.S., Thanki, N., Wang, Z., Yamashita, R.A., Zhang, D., Zheng, C., Bryant, S.H. 2015. CDD: NCBI's conserved domain database. Nucleic Acids Research, Vol. 43, Database Issue D222-2.

Margalef, R. 1978. Life-forms of phytoplankton as survival alternatives in an unstable environment. Oceanologica Acta, Vol. 1, pp. 493-509.

Markert, S. Arndt, C., Felbeck, H., Becher, D., Sievert, S.M., Hugler, M., Albrecht, D., Robidart, J., Bench, S., Feldman, R.A., Hecker, M., Schweder, T. 2007. Physiological proteomics of the uncultured endosymbiont of *Riftia pachyptila*. Science, Vol. 315, Issue 5809, pp. 247-50.

Martin, J.A., Wang, Z., 2011. Next-generation transcriptome assembly. Nature Review Genetics, Vol. 12, 671e682.

Martins, C. A., Kulis, D., Franca, S., Anderson, D.M. 2004. The loss of PSP toxin production in a formerly toxic *Alexandrium lusitanicum* clone. Toxicon, Vol. 43, pp. 195–205.

Massana, R., Gobet, A., Audic, S., Bass, D., Bittner, L., Boutte, C. et al. 2015. Marine protist diversity in European coastal waters and sediments as revealed by high-throughput sequencing. Environmental Microbiology, Vol. 17, pp. 4035-4049.

Matz, J.M., Blake, M.J., Tatelman, H.M., Lavoi, K.P., Holbrook, N.J. 1995. Characterization and regulation of cold-induced heat shock protein expression in mouse brown adipose tissue. American Journal of Physiology, Vol.269, No.1, p.38–47.

Maumus, F., Allen, A., Mhiri, C., Hu, H., Jabbari, K., Vardi, A., Grandbastien, M. and Bowler, C. 2009. Potential impact of stress activated retrotransposons on genome evolution in a marine diatom. BioMed Central Genomics, Vol. 10, Issue 624.

McClintock, B. 1951. Chromosome organization and genic expression. Cold Spring Harbor Symposium on Quantitative Biology, Vol. 16, pp. 13–47.

McDonald, S.M., Sarno, D., Scanlan, D.J. and Zingone, A. 2007a. Genetic diversity of eukaryotic ultraphytoplankton in the Gulf of Naples during an annual cycle. Aquatic Microbial Ecology, Vol. 50, pp. 75–89.

McDonald, S.M., Sarno, D., and Zingone, A. 2007b. Identifying *Pseudo-nitzschia* species in natural samples using genus-specific PCR primers and clone libraries. Harmful Algae, Vol. 6, pp. 849–860.

Menden-Deuer, S. and Lessard, E.J. 2000. Carbon to volume relationships for dinoflagellates, diatoms, and other protist plankton. Limnology and Oceanography, Vol. 45, pp. 569-579.

Menden-Deuer, S., Rowlett, J. 2014. Many ways to stay in the game: individual variability maintains high biodiversity in planktonic microorganisms. Journal of the Royal Society Interface, Vol. 11, Issue 95, 20140031.

Menzel, D.W. and Spaeth, J.P. 1962. Occurrence of vitamin B_{12} in the Sargasso Sea, Limnology and Oceanography, Vol. 7, pp. 151-154.

Meyer, F., Paarmann, D., D'Souza, M., Olson, R., Glass, E.M., et al. 2008. The metagenomics RAST server – a public resource for the automatic phylogenetic and functional analysis of metagenomes. BioMed Central Bioinformatics, Vol. 9, Issue 386.

Meyer, J. R., Ellner, S. P., Hairston, N. G. Jr, Jones, L.E., Yoshida, T. 2006. Prey evolution on the time scale of predator-prey dynamics revealed by allelespecific quantitative PCR. Proceedings of the National Academy of Sciences of U.S.A., Vol. 103, pp. 10690–10695.

Migicovsky, Z. and Kovalchuk, I. 2015. Transgenerational inheritance of epigenetic response to cold in *Arabidopsis thaliana.* Biocatalysis and Agricultural Biotechnology, Vol. 4, pp.1-10.

Minchin, P.R. 1987. An evaluation of relative robustness of techniques for ecological ordinations. Vegetatio, Vol. 69, Issue 1, pp. 89 -107.

Miura, K. and Furumoto, T. 2013. Cold signaling and cold response in plants. International Journal of Molecular Sciences, Vol. 14, pp. 5312-5337.

Miyazaki, T., Tainaka, K., Togashi, T., Suzuki, T., Yoshimura, J. 2006. Spatial coexistence of phytoplankton species in ecological timescale. Population Ecology, Vol. 48, pp. 107-112.

Mlouka, A., Comte, K., Castets, A.-M., Bouchier, C. and de Marsac, N.T. 2004. The gas vesicle gene cluster from Microcystis aeruginosa and DNA rearrangements that lead to loss of cell buoyancy. Journal of Bacteriology. Vol.186, p.2355–2365.

Mock, T., and B. M. Kroon. 2002. Photosynthetic energy conversion under extreme conditions. I: important role of lipids as structural modulators and energy sink under N-limited growth in Antarctic sea ice diatoms. II: the significance of lipids under light limited growth in Antarctic sea ice diatoms. Phytochemistry, Vol. 61, pp. 41-60.

Mock, T. and Valentin, K. 2004. Photosynthesis and cold acclimation: molecular evidence from a polar diatom. Journal of Phycology, Vol. 40, pp. 732-741.

Mock, T. and Hoch, N. 2005. Long-term temperature acclimation of photosynthesis in steady-state cultures of the polar diatom *Fragilariopsis cylindrus.* Photosynthesis Research, Vol. 85, pp.307-17.

Mock, T., Samanta, M.P., Iverson, V., Berthiaume, C., Robison, M., et al. 2008. Whole-genome expression profiling of the marine diatom *Thalassiosira pseudonana* identifies genes involved in silicon bioprocesses. Proceedings of the National Academy of Sciences of U. S. A., Vol. 105, pp. 1579–1584.

Mock, T. and Kirkham, A. 2011. What can we learn from genomics approaches in marine ecology? From sequences to eco-systems biology! Marine Ecology, Vol. 33, pp. 131-148.

Montes-Hugo, M., Doney, S.C., Ducklow, H.W., Fraser, W., Martinson, D., Stammerjohn, S.E. and Schofield, O. 2009. Recent changes in phytoplankton communities associated with rapid regional climate change along the Western Antarctic Peninsula. Science, Vol. 323, pp. 1470-1473.

Moon-van der Staay, S.Y., De Wachter, R., and Vaulot, D. 2001. Oceanic 18S rDNA sequences from picoplankton reveal unsuspected eukaryotic diversity. Nature, Vol. 409, pp. 607-610.

Moore, L. R., Rocap, G. and Chisholm, S. W. 1998. Physiology and molecular phylogeny of coexisting *Prochlorococcus* ecotypes. Nature, Vol. 393, pp. 464–7.

Morey, J.S., Monroe, E.A., Kinney, A.L., Beal, M., Johnson, J.G., Hitchcock, G.L., Van Dolah, F.M. 2011. Transcriptomic response of the red tide dinoflagellate, *Karenia brevis*, to nitrogen and phosphorus depletion and addition. BioMed Central Genomics, Vol. 12, Issue 1, pp. 346

Morin, P.A., Luikart, G., Wayne, R.K. 2004. SNP workshop group: SNPs in ecology, evolution and conservation. Trends in Ecology and Evolution, Vol.19, Issue 4, pp. 208-216.

Moustafa, A., Beszteri, B., Maier, U.G., Bowler, C., Valentin, K., Bhattacharya, D. 2009. Genomic footprints of a cryptic plastid endosymbiosis in diatoms. Science, Vol. 324, pp. 1724–1726.

Musacchia, F., Basu, S., Petrosino, G., Salvemini, M. and Sanges, R. 2015. Annocript: a flexible pipeline for the annotation of transcriptomes also able to identify putative long noncoding RNAs. Bioinformatics, Vol. 31, Issue 13, pp. 2199 - 201.

Naito, K., Zhang, F., Tsukiyama, T. et al. 2009. Unexpected consequences of a sudden and massive transposon amplification on rice gene expression. Nature, Vol. 461, pp. 1130–1134.

Nanjappa, D. 2012. Genetic, physiological and ecological diversity of the diatom genus *Leptocylindrus*. PhD Thesis, Open University, London, U.K. – Stazione Zoologica Anton Dohrn, Naples, Italy.

Nanjappa, D., Kooistra, W.H.C.F. and Zingone, A. 2013. A reappraisal of the genus *Leptocylindrus* (Bacillariophyta), with the addition of three species and the erection of *Tenuicylindrus* gen. nov. Journal of Phycology, Vol. 49, pp. 917-936.

Nanjappa, D., Audic, S., Romac, S., Kooistra, W.H.C.F., Zingone, A. 2014a. Assessment of species diversity and distribution of an ancient diatom lineage using a DNA metabarcoding approach. PloS One, Vol. 9, Issue 8, e103810.

Nanjappa, D., d'Ippolito, G., Gallo, C., Zingone, A. and Fontana, A. 2014b. Oxylipin diversity in the diatom family Leptocylindraceae reveals DHA derivatives in marine diatoms. Marine Drugs, Vol. 12, pp. 368-384.

Nanjappa, D., Sanges, R., Ferrante, M.I., Zingone, A. Submitted. Flagellar gene expression in a centric diatom during sexual reproduction. BioMed Central Genomics

Naydenov, M., Baev, V., Apostolova, E., Gospodinova, N., Sablok, et al. 2015. High-temperature effect on genes engaged in DNA methylation and affected by DNA methylation in *Arabidopsis.* Plant Physiology and Biochemistry, Vol. 87, pp.102-108.

Nelson, D.M., Tréguer, P., Brzezinski, A., Leynaert, A. and Quéguiner, B. 1995. Production and dissolution of biogenic silica in the ocean: revised global estimates, comparison with

regional data and relationship to biogenic sedimentation. Global Biogeochemical Cycles, Vol. 9, pp. 359-372.

Nelson, V.R., Heaney, J.D., Tesar, P.J., Davidson, N.O., et al. 2012. Transgenerational epigenetic efects of the Apobec1 cytidine deaminase deficiency on testicular germ cell tumor susceptibility and embryonic viability. Proceedings of the National Academy of Sciences USA, Vol. 109, pp. E2766–2773.

Nishizawa-Yokoi, A., Nosaka, R., Hayashi, H., Tainaka, H., Maruta, T., et al. 2011. HsfA1d and HsfA1e involved in the transcriptional regulation of *HsfA2* function as key regulators for the Hsf signaling network in response to environmental stress, Plant and Cell Physiology, Vol.52, Issue 5, pp.933-945.

Norden-Krichmar, T.M., Allen, A.E., Gaasterland, T. and Hildebrand, M. 2011. Characterization of the small RNA transcriptome of the diatom, *Thalassiosira pseudonana*. PLoS One, Vol. 6, e22870.

O'Brien, W. J. 1974. The dynamics of nutrient limitation of phytoplankton algae: A model reconsidered. Ecology, Vol. 55, pp. 135-141.

Odum, E. 1977. The emergence of ecology as a new integrative discipline. Science, Vol. 195, Issue 4284, pp. 1289-1293.

Pace, N. 2009. Mapping the tree of life: Progress and prospects. Microbiology and Molecular Biology Reviews, Vol. 73, Issue 4, pp. 565-576.

Pagel, M. 1994. The adaptationist wager. In: Eggleton, P. and Vane-Wright, R. (eds), Phylogenetics and ecology. Academic Press, London, pp. 29-51.

Palenik, B., Grimwood, J., Aerts, A., Rouze, A., Rouze, P., Salamov, A., Putnam, N., Dupont, C., Jorgensen R, et al. 2007. The tiny eukaryote *Ostreococcus* provides genomic insights into the paradox of plankton speciation. PNAS, Vol. 104, No.18, pp. 7705 – 7710.

Pallen, M.J., Matzke, N.J. 2006. From The Origin of Species to the origin of bacterial flagella. Nature Reviews Microbiology, Vol. 4, Issue 10, pp.784-90.

Palma-Guerrero, J., Torriani, S.F.F., Zala, M., Carter, D., Courbot, M., et al. 2016. Comparative transcriptomic analyses of *Zymoseptoria tritici* strains show complex lifestyle transitions and intraspecific variability in transcription profiles. Molecular Plant Pathology, Vol.17, Issue 6, pp. 845-859.

Parker, M.S. and Armbrust, E.V. 2005. Synergistic effects of light, temperature, and nitrogen source on transcription of genes for carbon and nitrogen metabolism in the centric diatom *Thalassiosira pseudonana* (Bacillariophyceae). Journal of Phycology, Vol.41, Issue 6, pp. 1142-1153.

Parker, M.S., Mock, T. and Armbrust, E.V. 2008. Genomic insights into marine microalgae. Annual Review of Genetics, Vol. 42, pp. 619–645.

Pedròs-Aliò, C. 2007. ECOLOGY: dipping into the rare biosphere. Science, Vol. 315, pp. 192–193.

Peers, G.,and Price, N.M. 2006. Copper-containing plastocyanin used for electron transport by an oceanic diatom. Nature, Vol. 441, pp. 341–344.

Pfaffl, M.W., Horgan, G. W., and Dempfle, L. 2002. Relative expression software tool (REST©) for group-wise comparison and statistical analysis of relative expression results in real-time PCR. Nucleic Acids Research, Vol.30, Issue 9, e36.

Pfeifer, F. and Blaseio, U. 1990. Transposition burst of the ISH27 insertion element family in *Halobacterium halobium*. Nucleic Acids Research, Vol.18, p.6921–6925.

Pigliucci, M. 2005. Evolution of phenotypic plasticity: where are we going now? Trends in Ecology and Evolution, Vol. 20, Issue 9, pp. 481–486.

Pörtner, H.O. and Farrell, A.P. 2008. Physiology and Climate Change. Science, Vol. 322, pp. 690-692.

Pride, D. T., Meinersmann, R. J., Wassenaar, T. M. and Blaser, M. J. 2003. Evolutionary implications of microbial genome tetranucleotide frequency biases. Genome Research, Vol. 13, pp. 145–158.

Putman, R.J. and Wratten, S.D. 1984. Principles of Ecology. University of California Press, Berkeley, California, USA.

Qian, J., Chen, J., Liu, Y.F., Yang, L.L., Li, W.P. and Zhang, L.M. 2014. Overexpression of *Arabidopsis HsfA1a* enhances diverse stress tolerance by promoting stress-induced *Hsp* expression. Genetics and Molecular Research, Vol.13, Issue 1, p.1233-1243.

Quast, C., Pruesse, E., Yilmaz, P., Gerken, J., Schweer, T., Yarza, P., Peplies, J., Glöckner, F.O. 2013. The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. Nucleic Acids Research, Vol. 41, Issue D1, pp. D590-D596.

Quigg, A., Finkel, Z.V., Irwin, A.J., Rosenthal, Y., Ho, T.Y., Reinfelder, J.R., Schofield, O., Morel, F.M.M., Falkowski, P.G. 2003. The evolutionary inheritance of elemental stoichiometry in marine phytoplankton. Nature, Vol. 425, pp. 291-294.

Ragazzola, F., Foster, L.C., Form, A.U., Buscher, J., Hansteen, T.H. and Fietzke, J. 2013. Phenotypic plasticity of coralline algae in a high CO_2 world. Ecology and Evolution, Vol. 3, Issue 10, pp. 3436-3446.

Rakocevic, A., Mondy, S., Tirichine, L., Cosson, V., Brocard, L., Iantcheva, A., Cayrel, A., Devier, B., El-Heba, G.A.A. and Ratet P. 2009. MERE1, a low-copy-number copia-type retroelement in Medicago truncatula active during tissue culture. Plant Physiology, Vol. 151, pp. 1250–1263.

Rakyan, V. K., Blewitt, M. E., Druker, R., Preis, J. I. and Whitelaw, E. 2002. Metastable epialleles in mammals. Trends in Genetics, Vol. 18, pp. 348–351.

Ralph, P. J., A. McMin, K. G. Ryan, and C. Ashworth. 2005. Short-term effect of temperature on the photokinetics of microalgae from the surface layers of Antarctic pack ice. Journal of Phycology, Vol. 41, pp.763–769.

Ramette, A. 2007. Multivariate analyses in microbial ecology. FEMS Microbiology Ecology, Vol.62, pp.142-160.

R Core Team. 2012 R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL: https://www.R-project.org/.

Rengefors, K. and Anderson, D.M. 1998. Environmental and endogenous regulation of cyst germination in two freshwater dinoflagellates. Journal of Phycology, Vol. 34, pp. 568-577.

Rengefors, K., Kremp, A., Reusch, T.B.H., Wood, M. 2017. Genetic diversity and evolution in eukaryotic phytoplankton: revelations from population genetic studies. Journal of Plankton Research, Vol. 39, Issue 2, pp. 165 – 179.

Reusch, T.B. and Boyd, P.W. 2013. Experimental evolution meets marine phytoplankton. Evolution, Vol.67, Issue 7, pp. 1849-1859.

Reva, O., Zaets, I., Ovcharenko, L., Kukharenko, O., Shpylova, S., Podolich, O., de Vera, J. and Kozyrovska. 2015. Metabarcoding of the kombucha microbial community grown in different microenvironments. AMB Express, Vol. 5, Issue 35.

Reynolds, C.S. 1998. What factors influence the species composition of phytoplankton in lakes of different trophic status? Hydrobiologia, Vol. 369, Issue 0, pp. 11-26.

Ribera D'Alcalà, M., Conversano, F., Corato, F., Licandro, P., Mangoni, O., et al. 2004. Seasonal patterns in plankton communities in a pluriannual time series at a coastal Mediterranean site (Gulf of Naples): an attempt to discern recurrences and trends. Scientia Marina, Vol. 68, pp. 65-83.

Richlen, M. L., Erdner, D.L., McCauley, L.A.R., Libera, K. and Anderson, D.M. 2012. Extensive genetic diversity and rapid population differentiation during blooms of *Alexandrium fundyense* (Dinophyceae) in an isolated salt pond on Cape Cod, MA, USA. Ecology and Evolution, Vol. 2, pp. 2583–2594.

Richmond, A. 1986. Cell response to environmental factors, In: Richmond A. (Ed.), Handbook of microalgal mass culture, CRC Press, Boca Raton, Florida.

Richter, K., Haslbeck, M., Buchner, J. 2010. The heat shock response: life on the verge of death. Molecular Cell, Vol.40, pp.253–266

Riesenfeld, C.S., Schloss, P.D., Handelsman, J. 2004. Metagenomics: genomic analysis of microbial communities. Annual Review of Genetics, Vol.38, pp.525–552.

Rines, J.E.B. and Hargraves, P.E. 1990. Morphology and taxonomy of Chaetoceros compressus Lauder var. hirtisetus var. nova, with preliminary consideration of closely related taxa. Diatom Research, Vol. 5, pp. 113-127.

Ritossa F. 1996. Discovery of the heat shock response. Cell Stress Chaperones, Vol.1, pp.97–98.

Robinson, M.D., McCarthy, D.J. and Smyth, G.K. 2010. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics, Vol. 26, pp. 139-140.

Rocap, G., Larimer, F.W., Lamerdin, J., Malfatti, S., Chain, P., et al. 2003. Genome divergence in two *Prochlorococcus* ecotypes reflects oceanic niche differentiation. Nature, Vol. 424, pp. 1042-1047.

Romero, P.R., Zaidi, S., Fang, Y.Y., Uversky, V.N., Radivojac, P., Oldfield, C.J., Cortese, M.S., Sickmeier, M., LeGall, T., Obradovic, Z., Dunker, A.K. 2006. Alternative splicing in concert with protein intrinsic disorder enables increased functional diversity in multicellular organisms. Proceedings of the National Academy of Sciences, Vol.103, pp. 8390–8395.

Ropars, J., Rodriguez de la Vega, R.C., López-Villavicencio, M., Gouzy, J., Sallet, E., et al. Adaptive horizontal gene transfers between multiple cheese-associated fungi. Current Biology, Vol. 25, Issue 19, pp. 2562-9.

Rose, J.M., Feng, Y., DiTullio, G.R., Dunbar, R.B., Hare, C.E., Lee, P.A., Lohan, M., Long, M., Smith Jr., W.O., Sohst, B., Tozzi, S., Zhang, Y. and Hutchins, D.A. 2009. Synergistic effects of iron and temperature on Antarctic phytoplankton and microzooplankton assemblages. Biogeosciences, Vol.6**,** pp. 3131 – 3147.

Ross, M.G., Russ, C., Costello, M., Hollinger, A., Lennon, N.J., Hegarty, R., Nusbaum, C. and Jaffe, D.B. 2013. Characterizing and measuring bias in sequence data. Genome Biology, Vol. 14, Issue 5, R51.

Round, F., Crawford, R. and Mann, D. 1990. The Diatoms: Biology and morphology of the genera. Cambridge University Press, Cambridge.

Rousch, J.M., Bingham, S.E., Sommerfeld, M.R. 2004. Protein expression during heat stress in thermos-intolerant and thermos-tolerant diatoms. Journal of Experimental Marine Biology and Ecology, Vol. 306, Issue 2, pp. 231-243.

Rowland, J.G., Pang, X., Suzuki, I., Murata, N., Simon, W.J., Slabas, A.R. 2010. Identification of components associated with thermal acclimation of photosystem II in *Synechocystis* sp. PCC6803, PLoS ONE, Vol. 5, Issue 5, e10511.

Roy, S., Beauchemin, M., Dagenais-Bellefeuille, S., Letourneau, L., Cappadocia, M., Morse, D. 2014. The *Lingulodinium* circadian system lacks rhythmic changes in transcript abundance. BioMed Central Biology, Vol. 12, Issue 107.

Rusch, D.B., Halpern, A.L., Sutton, G. et al. 2007. The Sorcerer II global ocean sampling expedition: northwest Atlantic through eastern tropical Pacific. PLoS Biology, Vol.5, Issue 3, e77.

Rynearson, T.A., Armbrust, E.V. 2004. Genetic differentiation among populations of the planktonic marine diatom *Ditylum brightwellii* (Bacillariophyceae). Journal of Phycology, Vol. 40, Issue 1, pp. 34–43.

Rynearson, T.A., Newton, J.A., Armbrust, E.V. 2006. Spring bloom development, genetic variation, and population succession in the planktonic diatom *Ditylum brightwellii*. Limnology and Oceanography, Vol. 51, pp. 1249–1261.

Saier, M. 2004. Evolution of bacterial type III protein secretion systems. Trends in Microbiology, Vol. 12, Issue 3, pp. 113–115.

Salvucci, M. E., and S. J. Crafts-Brandner. 2004. Relationship between the heat tolerance of photosynthesis and the thermal stability of rubisco activase in plants from contrasting thermal environments. Plant Physiology, Vol. 134, pp. 1460–1470.

Sangwan, V., Orvar, B.L., Beyerly, J., Hirt, H., Dhindsa, R.S. 2002. Opposite changes in membrane fluidity mimic cold and heat stress activation of distinct plant MAP kinase pathways. Plant Journal, Vol. 31, pp. 629–638.

Sarno, D., Kooistra, W.C.H.F., Medlin, L. K., Percopo, I. and Zingone, A. 2005. Diversity in the genus Skeletonema (Bacillariophyceae). II. An assessment of the taxonomy of *S. costatum*-like species, with the description of four new species. Journal of Phycology, Vol. 41, pp.151-176.

Sarno, D., Kooistra, W.C.H.F., Balzano, S., Hargraves, P.E. and Zingone, A. 2007. Diversity in the genus Skeletonema (Bacillariophyceae): III. Phylogenetic position and morphological variability of Skeletonema costatum and Skeletonema grevillei, with the description of Skeletonema ardens sp. nov. Journal of Phycology, Vol. 43, pp. 156-170.

Sarno, D. and Zingone, A. 2008. MARECHIARA-phytoplankton long-term time-series (1984-2006) at the fixed coastal station in the Gulf of Naples, Southern Tyrrhenian Sea. Stazione Zoologica Anton Dohrn.

Sarthou, G., TimmermanS, K.R., Blain, S. and Treguer, P. 2005. Growth physiology and fate of diatoms in the ocean: a review. Journal of Sea Research, Vol.53**,**pp.25-42.

Sayre, R. 2010. Microalgae: The Potential for Carbon Capture, Bioscience, Vol. 60, Issue 9, pp. 722-727.

Schaum, E., Rost, B., Millar, A.J., Collins, S. 2013. Variation in plastic responses of a globally distributed picoplankton species to ocean acidification. Nature Climate Change , Vol.3, Issue 3, pp. 298–302.

Scheffer, M., Rinaldi, S., Huisman, J., Weissing, F.J. 2003. Why plankton communities have no equilibrium: solutions to the paradox. Hydrobiologia, Vol. 491, Issue 1, pp. 9-18.

Schirmer, M., D'Amore, R., Ijaz, U.Z., Hall, N. and Quince, C. 2016. Illumina error profiles: resolving fine-scale variation in metagenomics sequencing data. BMC Bioinformatics, Vol. 17, Issue 125.

Schlesinger, M.J., Aliperti, G., Kelley, P.M., 1982. The response of cells to heat shock. Trends in Biochemical Sciences, Vol. 7, pp. 222– 225.

Schlichting, C., and M. Pigliucci. 1996. Phenotypic evolution: a reaction norm perspective. Heredity, Vol. 82, pp. 344.

Schloss, P.D., Westcott, S.L., Ryabin, T., Hall, J.R., Hartmann, M., Hollister, E.B. et al. 2009. Introducing mothur: Open-source, platform-independent, community-supported software

for describing and comparing microbial communities. Applied and Environmental Microbiology, Vol.75, Issue 23, pp. 7537-41.

Schlüter, L., Lohbeck, K.T., Gutowska, M.A., Gröger, J.P., Riebesell, U., Resusch, B.H. 2014. Adaptation of a globally important coccolithophore to ocean warming and acidification. Nature Climate Change, Vol. 4, pp. 1024 – 1030.

Schmid, A.M.M. 2003. Endobacteria in the diatom *Pinnularia* (Bacillariophyceae). I. Scattered ct-nucleoids explained: DAPI-DNA complexes stem from exoplastidial bacteria boring into the chloroplasts. Journal of Phycology, Vol. 39, pp. 122–38

Schmidt, R., Schippers, J.H.M., Welker, A., Mieulet, D., Guiderdoni, E. and Mueller-Roeber, B. 2012. Transcription factor OsHsfC1b regulates salt tolerance and development in *Oryza sativa* spp. *japonica*. AoB Plants, pls011.

Schmitt, E., Gehrmann, M., Brunet, M., Multhoff, G., Garrido, C. 2007. Intracellular and extracellular functions of heat shoch proteins: repercussions in cancer therapy. Journal of Leukocyte Biology, Vol.81, no.1, pp. 15-27.

Schuster, S.C. 2008. Next-generation sequencing transforms today's biology. Nature Methods, Vol. 5, Issue 1, pp. 16-18.

Schwender, J., Ohlrogge, J., Shachar-Hill, Y. 2004. Understanding flux in plant metabolic networks, Current Opinion in Plant Biology, Vol. 7, pp.309–317.

Scoccianti, V., Penna, A., Penna, N., Magnani, M., 1995. Effect of heat stress on polyamine content and protein pattern in *Skeletonema costatum*. Marine Biology, Vol. 121, pp. 549–554.

Sebens, K. and Thorne, B. 1985. Coexistence of clones, clonal diversity, and the effects of disturbance. In Jackson, J., Buss, L. and Cook, R. [Eds.] Population Biology and Evolution of Clonal Organisms. Yale University Press, New Haven, pp. 357–98.

Seiser, R. M., Sundberg, A. E., Wollam, B. J., Zobel –Thropp, P., Baldwin, K., Spector, M. D. and Lycan, D.E. 2006. Ltv1 is required for efficient nuclear export of the ribosomal small subunit in *Saccharomyces cerevisiae.* Genetics Society of America, Vol. 174, pp. 679 – 691.

Serra, R. 2008. Role of intraflagellar transport and primary cilia in skeletal development. The Anatomical Record, Vol. 291, pp. 1049 – 1061

Seyednasrollah, F., Laiho, A. and Elo, L.L. 2013. Comparison of software packages for detecting differential expression in RNA-seq studies. Briefings in Bioinformatics, Vol. 16, Issue 1, pp. 59 – 70.

Shamovsky, I. and Nudler, E. 2008. New insights into the mechanism of heat shock response activation. Cellular and Molecular Life Sciences, Vol.65, pp. 855-861.

Shendure, J., Hanlee, J. 2008. Next-generation DNA sequencing. Nature Biotechnology, Vol.26, pp. 1135-1145.

Shirokawa, Y., Karino, K., Mayama, S. 2012. Developmental plasticity and genotype-environment interactions influence valve morphology in the *Cyclotella meneghiniana* species complex (Bacillariophyceae). European Journal of Phycology, Vol. 47, pp. 245–253.

Shoemaker, R., Deng, J., Wang, W., Zhang, K. 2010. Allele-specific methylation is prevalent and is contributed by CpG-SNPs in the human genome. Genome Research, Vol. 20, pp. 883–889.

Siaut, M., Heijde, M., Mangogna, M., Montsant, A., Coesel, S., Allen, A., Manfredonia, A., Falciatore, A. and Bowler, C. 2007. Molecular toolbox for studying diatom biology in *Phaeodactylum tricornutum*. Gene, Vol. 406, pp. 23-35.

Sijen, T. and Plasterk, R. H. 2003. Transposon silencing in the *Caenorhabditis elegans* germ line by natural RNAi. Nature, Vol. 426, pp. 310–314.

Simola, D.F., Ye, C., Mutti, N.S., et al. 2013. A chromatin link to caste identity in the carpenter ant *Camponotus floridanus*. Genome Research, Vol. 23, pp.486–496.

Simon, N., Cras, A.-L., Foulon, E. and Lemee, R. 2009. Diversity and evolution of marine phytoplankton. Comptes Rendus Biologies, Vol. 332, pp. 159–170.

Simpson, J.T., Wong, K., Jackman, S.D., Schein, J.E., Jones, S.J., Birol, I. 2009. ABySS: A parallel assembler for short read sequence data. Genome Research, Vol.19, pp. 1117-1123.

Sims, P.A., Mann, D.G., and Medlin, L.K. 2006. Evolution of the diatoms: insights from fossil, biological and molecular data. Phycologia, Vol. 45, pp. 361–402.

Slotkin, R.K. and Martienssen, R. 2007. Transposable elements and the epigenetic regulation of the genome. Nature Reviews Genetics, Vol.8, pp.272-285.

Smayda, T. 1980. Phytoplankton species succession in the physiological ecology of phytoplankton. (Ed. by I. Morris), Blackwell Scientific Publications, Oxford, pp. 493-570.

Smetacek, V. 1999. Diatoms and the ocean carbon cycle. Protist, Vol. 150, pp. 25-32.

Smirnova, G.V., Zakirova, O.N., Oktiabr'skii, O.N. 2001. Role of the antioxidant system in response of *Escherichia coli* bacteria to cold stress. Mikrobiologiia, Vol. 70, pp.55–60.

Smit, A.F.A., Hubley, R. and Green, P. 2013-2015. RepeatMasker Open-4.0 <http://www.repeatmasker.org>

Smith-Unna, R.D., Boursnell, C., Patro, R., Hibberd, J.M., Kelly, S. TransRate: reference free quality assessment of de-novo transcriptome assemblies. Genome Research, Vol. 26, Issue 8, pp. 1134-44.

Snyder, L.A.S., Davies, J.K., Saunders, N.J. 2004. Microarray genomotyping of key experimental strains of *Neisseria gonorrhoeae* reveals gene complement diversity and five new neisserial genes associated with Minimal Mobile Elements. BioMed Central Genomics, Vol. 5, 23.

Sobkowiak, A., Jonczyk, M., Jarochowska, E., Biecek, P., Trzcinska-Danielewicz, J., et al. 2014. Genome‐wide transcriptomic analysis of response to low temperature reveals candidate genes determining divergent cold‐sensitivity of maize inbred lines. Plant Molecular Biology, Vol. 85, pp. 317-331.

Sogin, M.L., Morrison, H.G., Huber, J.A. et al. 2006. Microbial diversity in the deep sea and the underexplored ‘‘rare biosphere’’. Proceedings of the National Academy of Sciences USA, Vol. 103, pp. 12115–12120.

Somerville, C. 1995. Direct tests of the role of membrane lipid composition in low temperature-induce photoinhibition and chilling sensitivity in plants and cyanobacteria, Proceedings of the National Academy of Sciences (USA), Vol. 92, pp. 6215–6218.

Sommer, U. 1985. Seasonal succession of phytoplankton in Lake Constance. BioScience, Vol. 35, pp. 351-357.

Stearns, S. C. 1986. Natural selection and fitness, adaptation and constraint. In: Raup. D. M. and Jablonski. D. (eds), Patterns and processes in the history of life. Springer, Heidelberg, pp. 2344.

Stearns, S. C. 1989. The evolutionary significance of phenotypic plasticity – phenotypic sources of variation among organisms can be described by developmental switches and reaction norms. Bioscience, Vol. 39, pp. 436–45.

Stoebel, D.M., Dorman, C.J. 2010. The effect of mobile element IS10 on experimental regulatory evolution in *Escherichia coli*. Molecular Biology and Evolution, Vol.27, pp. 2105–2112.

Stoeck, T., Epstein, S. 2003. Novel eukaryotic lineages inferred from small-subunit rRNA analyses of oxygen-depleted marine environments. Applied and Environmental Microbiology, Vol. 69, pp. 2657–2663.

Stoeck, T., Hayward, B., Taylor, G.T., Varela, R., and Epstein, S.S. 2006. A multiple PCR-primer approach to access the microeukaryotic diversity in environmental samples. Protist, Vol. 157, pp. 31-43.

Stouder, C., Paoloni-Giacobino, A. 2011. Specific transgenerational imprinting effects of the endocrine disruptor methoxychlor on male gametes. Reproduction, Vol. 141, pp. 207–216.

Strzepek, R.F., Harrison, P.J. 2004. Photosynthetic architecture differs in coastal and oceanic diatoms. Nature, Vol. 431, pp. 689–692.

Stuart, R. K., Brahamsha, B., Busby, K. and Palenik, B. 2013. Genomic island genes in a coastal marine *Synechococcus* strain confer enhanced tolerance to copper and oxidative stress. The ISME Journal, Vol. 7, pp. 1139–49.

Studer, R.A., Robinson-Rechavi, M. 2009. How confident can we be that orthologs are similar, but paralogs differ? Trends in Genetics, Vol.25, pp. 210–216.

Sun, B., Zhu, Z., Cao, P., Chen, H., Chen, C., et al. 2016. Purple foliage coloration in tea (*Camellia sinensis L*.) arises from activation of the R2R3-MYB transcription factor CsAN1. Scientific Reports, Vol. 6, 32534.

Supek, F., Bošnjak, M., Škunca, N., Šmuc, T. 2011. REVIGO summarizes and visualizes long lists of Gene Ontology terms. PLoS ONE, Vol. 6, Issue 7, e21800.

Suzek, B.E., Huang, H., McGarvey, P., Mazumder, R. and Wu, C.H. 2007. UniRef: comprehensive and non-redundant UniProt reference clusters. Bioinformatics, Vol.23, Issue 10, pp.1282-1288.

Tabara, H. et al. 1999. The rde-1 gene, RNA interference, and transposon silencing in *C. elegans*. Cell, Vol. 99, pp. 123–132.

Taberlet, P., Coissac, E., Pompanon, F., Brochmann, C. and Willerslev, E. 2012. Towards next-generation biodiversity assessment using DNA metabarcoding. Molecular Ecology, Volume 21, pp. 2045-2050.

Tamura, K., Stecher, G., Peterson, D., Filipski, A., and Kumar, S. 2013. MEGA6: Molecular Evolutionary Genetics Analysis Version 6.0. Molecular Biology and Evolution, Vol. 30, pp. 2725-2729.

Thamatrakoln, K., Korenovska, O., Niheu, A.K., Bidle, K.D. 2012. Whole-genome expression analysis reveals a role for death-related genes in stress acclimation of the diatom *Thalassiosira pseudonana.* Environmental Microbiology, Vol. 14, Issue 1, pp. 67-81.

Thomashow, M.F. 1999. Plant cold acclimation: freezing tolerance genes and regulatory mechanisms, Annual Review of Plant Physiology, Vol. 50, pp. 571–599.

Thompson, A. W., Huang, K., Saito, M. A. and Chisholm, S. W. 2011. Transcriptome response of high- and low-light-adapted *Prochlorococcus* strains to changing iron availability. ISME Journal, Vol. 5, pp. 1580–1594.

Thompson, J. D. 1991. Phenotypic plasticity as a component of evolutionary change. Trends in Ecology and Evolution, Vol. 6, pp. 246–249.

Tilman, D. 1977. Resource competition between planktonic algae: an experimental and theoretical approach. Ecology, Vol. 58, pp. 338-348.

Tilman, D., Mattson, M. and Langer, S. 1981. Competition and nutrient kinetics along a temperature gradient: An experimental test of mechanistic approach to niche theory. Limnology and Oceanography, Vol. 26, Issue 6., pp. 1020-1033.

Timperio, A.M., Egidi, M.G., Zolla, L. 2008. Proteomics applied on plant abiotic stresses: Role of heat shock proteins (HSP). Journal of Proteomics, Vol. 71, pp. 391–411.

Tirichine, L., Rastogi, A., Bowler, C. 2017. Recent progress in diatom genomics and epigenomics. Current Opinion in Plant Biology, Vol. 36, pp. 46-55.

Toseland, A., Daines, S.J., Clark, J.R., Kirkham, A., Strauss, J., et al. 2013. The impact of temperature on marine phytoplankton resource allocation and metabolism. Nature Climate Change, Vol. 3, pp. 979 - 984.

Trott, A. and Morano, K.A. 2004. *SYM1* is the stress-induced *Saccharomyces cerevisiae* ortholog of the mammalian kidney disease gene *Mpv17* and is required for ethanol metabolism and tolerance during heat shock. Eukaryotic Cell, Vol.3, No.3, pp. 620-631.

Tunnicliffe, V. and Fowler, M. 1996. Influence of sea-floor spreading on the global hydrothermal vent fauna. Nature, Vol. 379, pp. 531-533.

Ullrich, S., Kube, M., Schubbe, S., Reinhardt, R. and Schuler, D. 2005. A hypervariable 130-kilobase genomic region of *Magnetospirillum gryphiswaldense* comprises a magnetosome island which undergoes frequent rearrangements during stationary growth. American Society for Microbiology, Vol.187, No.21, pp. 7176-7184.

Van Dover, C.L. 2000. The ecology of deep-sea hydrothermal vents. Princeton University Press, New Jersey, USA.

Vanelslander, B., Creach, V., Vanormelingen, P., Ernst, A., Chepurnov, V.A., Sahan, E., Muyzer, G., Stal, L.J., Vyverman, W., Sabbe, K., 2009. Ecological differentiation between sympatric pseudocryptic species in the estuarine benthic diatom *Navicula phyllepta* Bacillariophyceae. Journal of Phycology, Vol 45, pp. 1278–1289.

Van Leeuwen, P. J., de Ruijter, W. P. M., Lutjeharms, J. R. E. 2000. Natal pulses and the formation of Agulhas rings. Journal of Geophysical Research, Vol. 105 (C3), pp. 6425–6436.

Vanlerberghe, G.C. 2013. Alternative Oxidase: a mitochodrial respiratory pathway to maintain metabolic and signaling homeostasis during abiotic and biotic stress in plants. International Journal of Molecular Sciences, Vol. 14, pp. 6805-6847.

Vastenhouw, N.L., Brunschwig, K., Okihara, K.L., Muller, F., et al. 2006. Gene expression: long-term gene silencing by RNAi. Nature, Vol. 442, Issue 7105, p. 882.

Veluchamy, A., Lin, X., Maumus, F., Rivarola, M., Bhavsar, J., et al. 2013. Insights into the role of DNA methylation in diatoms by genome-wide profiling in *Phaeodactylum tricornutum*. Nature Communications, Vol. 4, 2091.

Vertii, A., Alison, B., Delaval, B., Hehnly, H., Doxsey, S. 2015. New frontiers: discovering cilia-independent functions of cilia proteins. EMBO reports, Vol. 16, pp. 1275-1287.

Vieira, C., Aubry, P., Lepetit, D. and Biemont, C. 1998. A temperature cline in copy number for 412 but not roo/B104 retrotransposons in populations of *Drosophila simulans*. Proceedings of the Royal Society B: Biological Sciences, Vol. 265, pp. 1161–1165.

Villar, E., Farrant, G.K., Follows, M., Garczarek, L., Speich, S. et al. 2015. Environmental characteristics of Agulhas rings affect interocean plankton transport. Science, Vol. 348, Issue 6237.

von Dassow, P., Ogata, H., Probert, I., Wincker, P., Da Silva, C., Audic, S., Claverie, J., de Vargas, C. 2009. Transcriptome analysis of functional differentiation between haploid and diploid cells of *Emiliania huxleyi*, a globally significant photosynthetic calcifying cell. Genome Biology, Vo.10, Issue 10, Article R114.

Walsh, J.J. 1976. Herbivory as a factor in patterns of nutrient utilization in the sea. Limnology and Oceanography, Vol. 21, pp. 1-13.

Wang, Q., Garrity, G.M., Tiedje, J.M., and Cole, J.R. 2007. Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. Applied and Environmental Microbiology, Vol. 73, pp. 5261-5267.

Wanntorp, H.-E. 1983. Historical constraints in adaptation theory: traits and non-traits. Oikos, Vol. 41, pp. 157-159.

Warnecke, F., Hugenholtz, P. 2007. Building on basic metagenomics with complementary technologies. Genome Biology, Vol. 8, Issue 12, pp. 231.

Watts, P.C., Martin, L.E., Kimmance, S.A., Montagnes, D.J.S., Lowe, C.D. 2011. The distribution of *Oxyrrhis marina*: a global disperser or poorly characterized endemic? Journal of Plankton Research, Vol. 33, Issue 4, pp. 579–589.

Weisse, T. 2008. Distribution and diversity of aquatic protists: an evolutionary and ecological perspective. Biodiversity and Conservation, Vol. 17, Issue 2, pp. 243–259.

Welti, R., Li, W., Li, M., Sang, Y., Biesiada, H., Zhou, H.E., Rajashekar, C.B., Williams, T.D., Wang, X. 2002. Profiling membrane lipids in plant stress responses. Role of phospholipase D alpha in freezing-induced lipid changes in *Arabidopsis*. The Journal of Biological Chemistry, Vol. 30, Issue 227, pp.31994-32002

West-Eberhard, M.J. 2003. Developmental plasticity and evolution. Oxford, UK: Oxford University Press.

Whittaker, K., Rignanese, D., Olson, R., Rynearson, T. 2012. Molecular subdivision of the marine diatom Thalassiosira rotula and its relationship to differences in geographic distribution, genome size, and physiology. BMC Evolutionary Biology, Vol. 12, pp. 209–217.

Wicker, T., Sabot, F., Hua-Van, A., Bennetzen, J., Capy, P., et al. 2007. A unified classification system for eukaryotic transposable elements. Nature Reviews Genetics, Vol. 8, pp. 973–982

Wintzingerode, F.V., Göbel, U.B., and Stackebrandt, E. 1997. Determination of microbial diversity in environmental samples: pitfalls of PCR-based rRNA analysis. FEMS Microbiology Reviews, Vol. 21, pp. 213-229

Woese, C. R. 1987. Bacterial evolution. Microbiological Reviews, Vol. 51, pp. 221–271.

Wohlrab, S., Tillmann, U., Cembella, A. and John, U. 2016. Trait changes induced by species interactions in two phenotypically distinct strains of a marine dinoflagellate. International Society for Microbial Ecology, Vol. 10, Issue 11, pp. 2658-2668.

Wolf, J.B.W. 2013. Principles of transcriptome analysis and gene expression quantification: an RNA-seq tutorial. Molecular Ecology Resources, Vol. 13, pp. 559-572.

Wong, K. M., Suchard, M. A. and Huelsenbeck, J. P. 2008. Alignment uncertainty and genomic analysis. Science, Vol. 319, pp. 473–476.

Wood, A.M., Leatham, T. 1992. The species concept in phytoplankton. Journal of Phycology, Vol. 28, Issue 6, pp. 723 – 729.

Wood, A.M., Everroad, R. C. and Wingard, L. M. 2005. Measuring algal growth rates in microalgal cultures. In: Anderson, R. A. Algal Culturing Techniques. Elsevier Academic Press, Burlington, pp. 269–285.

Worden, A.Z., Lee, J.H., Mock, T., Rouze, P., Simmons, M.P., et al. 2009. Green evolution and dynamic adaptations revealed by genomes of the marine picoeukaryotes Micromonas. Scienece, Vol. 324, Issue 5924, pp. 268-72.

Wu, C. 1995. Heat shock transcription factors: structure and regulation. Annual Review of Cell and Developmental Biology, Vol.11, pp. 441– 469.

Wu, M.T., Chatterji, S. and Eisen, J.A. 2012. Accounting for alignment uncertainty in phylogenomics. PLoS ONE, Vol. 7, e30288.

Wunderlich, M., Groß‐Hardt, R., Schöff, F. 2014. Heat shock HSFB2a involved in gametophyte development of *Arabidopsis thaliana* and its expression is controlled by a heat-inducible long non-coding antisense RNA. Plant Molecular Biology, Vol.85, pp. 541-550.

Xiong, Y. and Eickbush, T. 1990. Origin and evolution of retroelements based upon their reverse transcriptase sequences. The EMBO Journal, Vol. 9, No.10, pp. 3353-3362.

Yamaguchi-Shinozaki, K., Shinozaki, K. 2006. Transcriptional regulatory networks in cellular responses and tolerance to dehydration and cold stresses, Annual Review of Plant Biology, Vol. 57, pp. 781–803.

Yamazaki, S., Nomata, J. and Fujita, Y. 2006. Differential operation of dual protochlorophyllide reductases for chlorophyll biosynthesis in response to environmental oxygen levels in the cyanobacterium *Leptolyngbya boryana.* Plant Physiology, Vol. 142, pp. 911-922.

Yinqiu, J., Ashton, L., Pedley, S.M., Edwards, D.P., Tang, Y., et al. 2013. Reliable, verifiable and efficient monitoring of biodiversity via metabarcoding. Ecology Letters, Vol.16, pp. 1245-1257.

Yoshida, T., Hairston, N. G. Jr and Ellner, S. P. 2004. Evolutionary trade-off between defense against grazing and competitive ability in a simple unicellular alga, *Chlorella vulgaris*. Proceedings of the Royal Society of London Series B, Vol. 271, pp. 1947–1953.

Zarraonaindia, I., Smith, D.P., Gilbert, J.A. 2013. Beyond the genome: community level analysis of the microbial world. Biological Philosophy, Vol. 28, pp. 261-282.

Zhan, A., Hulak, M., Sylvester, F., Huang, X., Adebayo, A.A., et al. 2013. High sensitivity of 454 pyrosequencing for detection of rare species in aquatic communities. Methods in Ecology and Evolution, Vol. 4, pp. 558–565.

Zhan, A., Xiong, W., He, S., Maclsaac, H.J. 2014. Influence of artifact removal on rare species recovery in natural complex communitites using high-throughput sequencing. PLoS ONE, Vol. 9, Issue 5, e96928.

Zheng, W., and S., Kathariou. 1994. Transposon-induced mutants of *Listeria monocytogenes* incapable of growth at low temperature (4°C). FEMS Microbiology Letters, Vol. 121, pp. 287–292.

Zhu, X., Thalor, S.K., Takahashi, Y., Berberich, T., Kusano, T. 2012. An inhibitory effect of the sequence-conserved upstream open-reading frame on the translation of the main open-reading frame of HsfB1 transcripts in *Arabidopsis.* Plant, Cell and Environment, Vol. 35, Issue 11, pp. 2014-30.

Zhu, Y., Dai, J., Fuerst, P.G., Voytas, D.F. 2003. Controlling integration specificity of a yeast retrotransposon. Proceedings of the National Academy of Sciences, USA, Vol. 100, pp. 5891–5895.

Zimmermann, J., Glöckner, G., Jahn, R., Enke, N., and Gemeinholzer, B. 2015. Metabarcoding vs. morphological identification to assess diatom diversity in environmental studies. Molecular Ecology Resources, Vol. 15, pp. 526-542.

Zinger, L., Gobet, A., and Pommier, T. 2012. Two decades of describing the unseen majority of aquatic microbial diversity. Molecular Ecology, Vol. 21, pp. 1878-1896.

Zingone, A., Casotti, R., Ribera D'Alcalà, M., Scardi, M. and Marino, D. 1995. 'St Martin's Summer': the case of an autumn phytoplankton bloom in the Gulf of Naples (Mediterranean Sea). Journal of Plankton Research, Vol. 17, pp. 575-593.

Zingone, A., Percopo, I., Sims, P. A., Sarno, D. 2005. Diversity in the genus *Skeletonema* (Bacillariophyceae): I. A reexamination of the type material of *S. costatum* with the description of *S. grevillei* sp. nov. Journal of Phycology, Vol. 41, pp. 140-150.

Zovoilis, A., Cifuentes-Rojas, C., Chu, H., Hernandez, A.J., Lee, J.T. 2016. Destabilization of B2 RNA by EZH2 activates the stress response. Cell, Vol. 167, pp. 1788-1802.

8. Appendices

Appendix 1

Classes of transposable elements

Transposable elements can be divided into two classes based on the main mechanism of movement and their transposition intermediate (Wicker et al., 2007):

1. Class I TEs or "retrotransposons". These TEs follow the 'copy-and-paste' replicative mechanism where a reverse transcription process results in one or several additional copies of the original element that are subsequently inserted in other loci. Due to this kind of mechanism, this class of TEs can rapidly invade a genome. Retrotransposons can be further divided into five orders depending on the organization of their gene domain and the presence or absence of small repeats of non-coding domains: (a) long term repeat (LTR) order, (b) long interspersed repetitive elements (LINEs) order, (c) *Dictyostelium* intermediate repeat sequences (DIRS) order, (d) Penelope (PLE) order and (e) short interspersed repetitive elements (SINEs) order (Fig. A1.1).

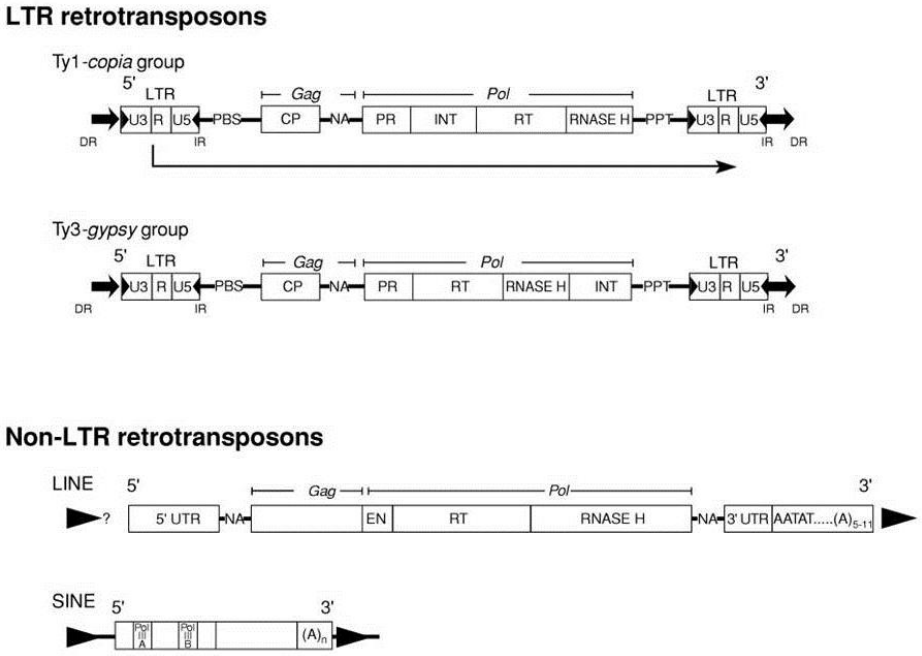


Figure A1.1 General structures of the Ty1-Copia, Ty3-gypsy which are LTR retrotransposons, LINE and SINE retrotransposons in plants. The LTR retrotransposons have long terminal repeats (LTR) in direct orientation at each end. Within the LTRs are U3, R, and U5 regions that contain signals for initiation and termination of transcription. The transcript is represented by the thin arrow. The genes within the retrotransposons encode capsid-like proteins (CP), endonuclease (EN), integrase (INT), protease (PR), reverse transcriptase (RT), and RNAase-H. Other sequences featured are PBS (primer binding sites), PPT (polypurine tracts), NA (nucleic acid binding moiety), IR (inverted terminal repeats), DR (flanking target direct repeat), 5' UTR (5' untranslated region), 3' UTR (3' untranslated region), and Pol III A and B-promoter recognition sites for RNA polymerase III. The envelope (env) gene-like sequence in the position of ORF 3, where a functional env gene is present in the animal retroviruses, has been found in both Ty1-copia and Ty3-gypsy groups (not shown). The LTR retrotransposons range from a few kb up to 15 kb in size. LINEs usually range in size from less than 1 kb to ca 8 kb, while SINEs are normally 100 bp to 300 bp in size. (Kumar and Bennetzen, 1999).

2. Class II TEs or “DNA transposons” which follow the ‘cut-and-paste’ mechanism. These TEs are transposed via a DNA intermediate; the original element is excised and reinserted at another locus of the genome. This mechanism is more conservative and less genome-invading than the ‘copy-and-paste’ of the retrotransposons. However, some DNA transposons (subclass 2) use the displacement of only one DNA strand getting out of the ‘cut-and-paste’ boundaries. Class II consists of four order: (a) TIR order, (b) Crypton, (c) Helitron, and (d) Maverick order.

Appendix 2

World maps in DCM Tara samples

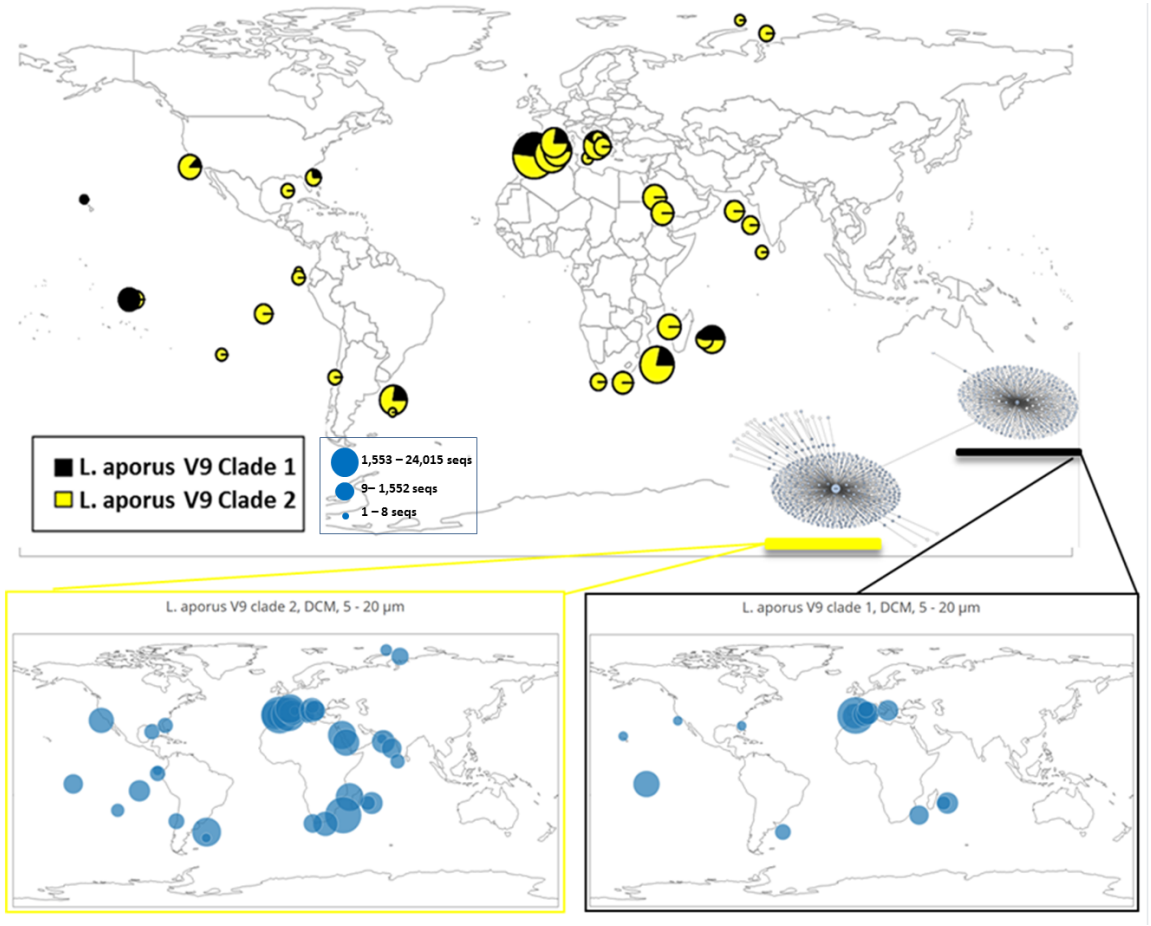


Figure A2.1. World distribution of log (abundance+1) of *L. aporus* at the Tara stations’ at the DCM samples, 5-20 µm size fraction. OTUs were represented by swarms (Mahè et al., 2014) and were linked to the spatial distribution with corresponding colours.

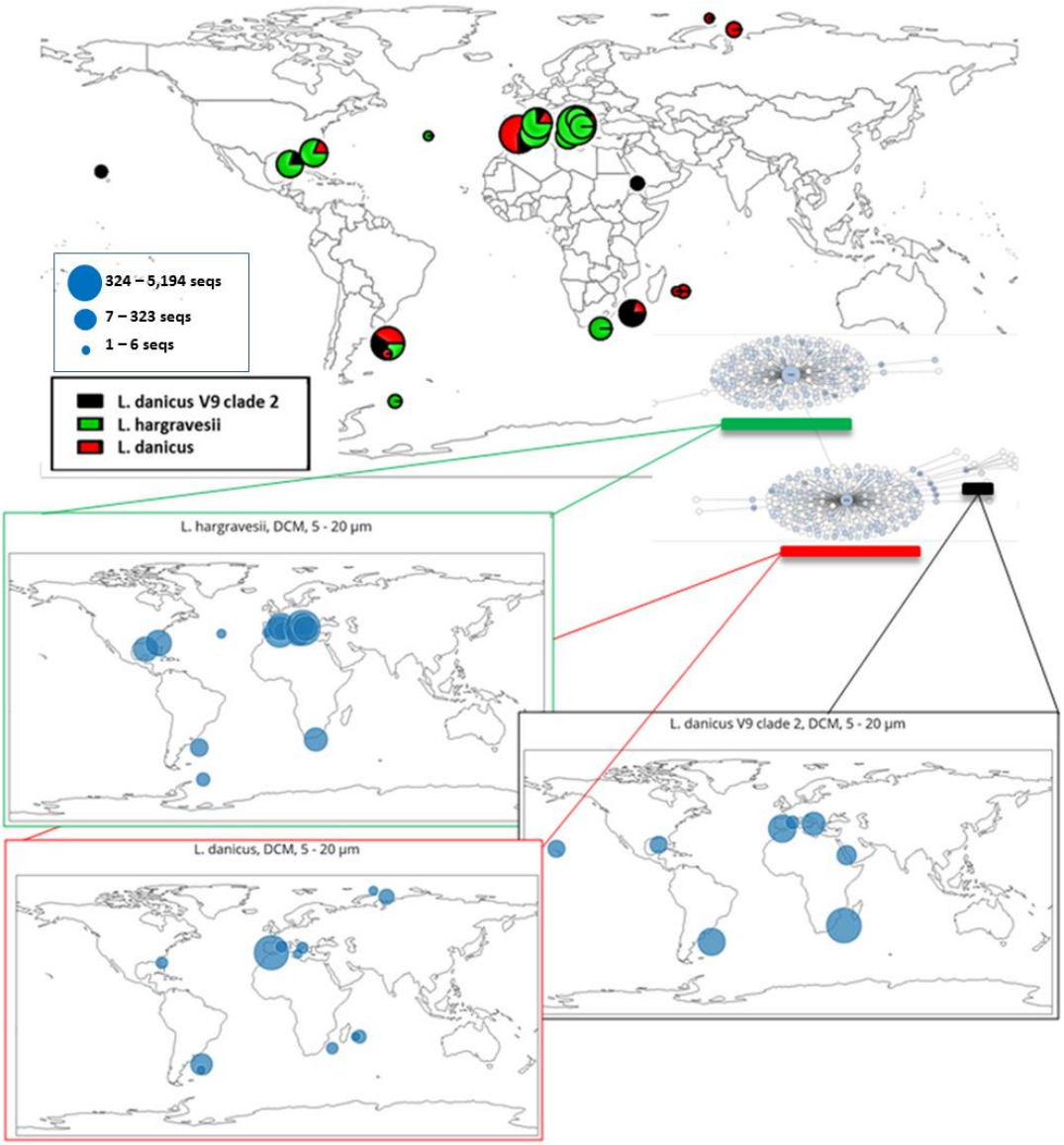


Figure A2.2 World distribution of log (abundance+1) of *L. danicus* clades and *L. hargravesii* at the Tara stations' at the DCM samples, 5-20 µm size fraction. OTUs were represented by swarms (Mahè et al., 2014) and were linked to the spatial distribution with corresponding colours.

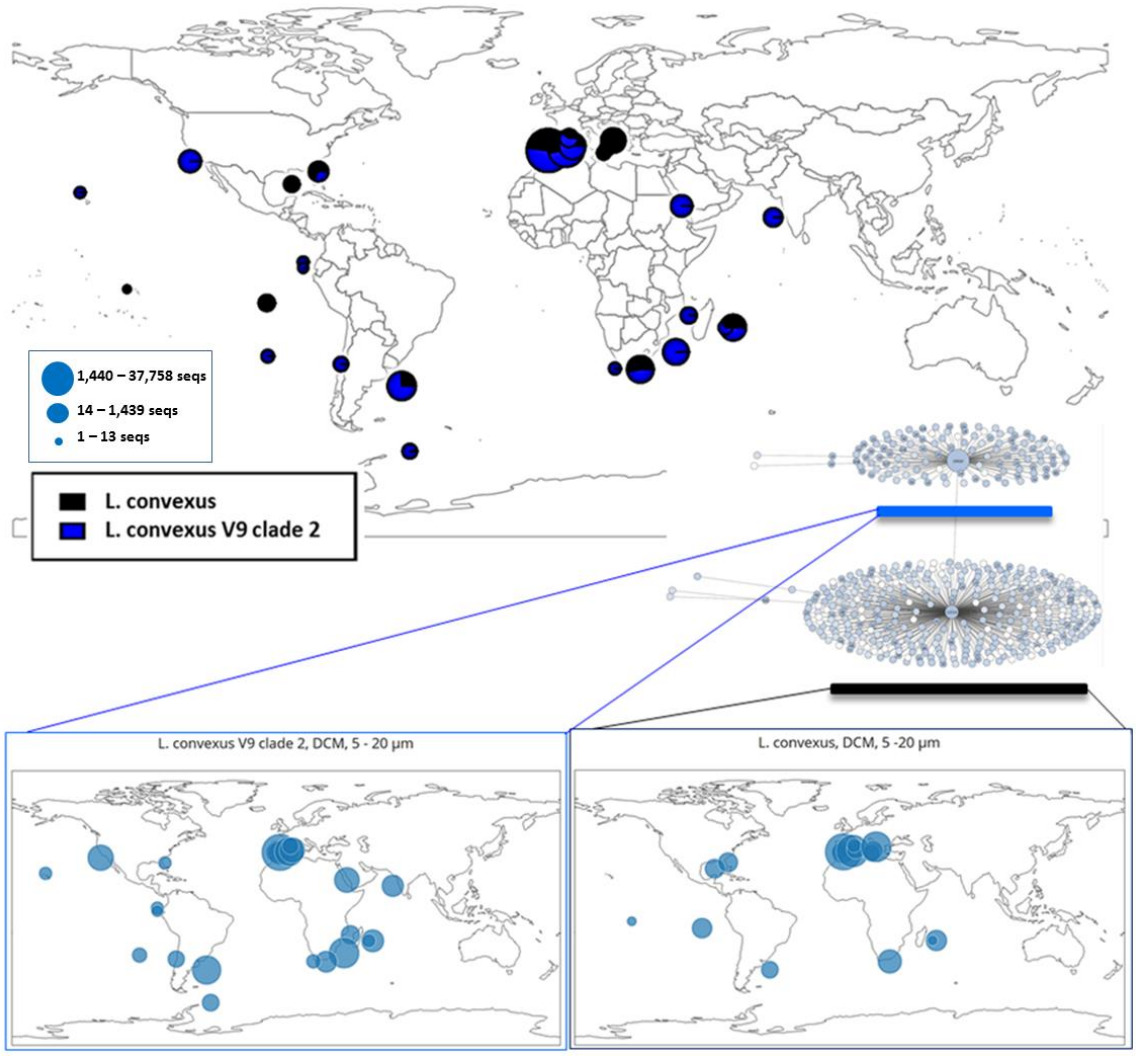


Figure A2.3. World distribution of log (abundance+1) of *L. convexus* at the Tara stations' DCM samples, 5-20 µm size fraction. OTUs were represented by swarms (Mahè et al., 2014) and were linked to the spatial distribution with corresponding colours.

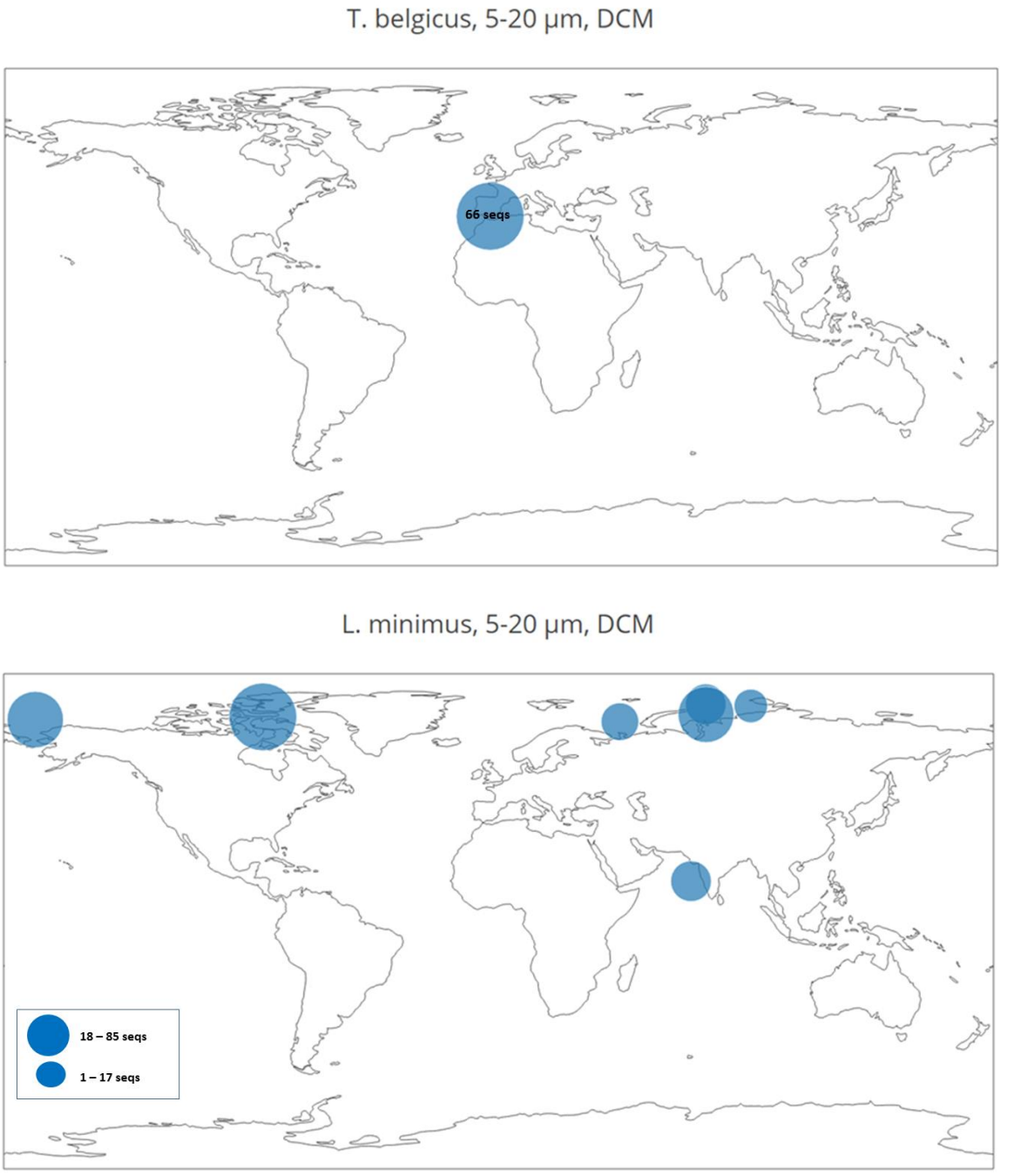


Figure A2.4. World distribution of *L. minimus* and *T. belgicus* separately at Tara DCM samples, 5 -20 µm size fraction. The size of the bubbles represents the normalized abundance, log (abundance+1), within each clade.